NATURE INSIGHT MEMBRANE PROTEIN BIOPHYSICS

9 May 2007 www.nature.com/nature £10

nature

DIGITAL STORAGE IN FIVE DIMENSIONS

How to cram 1.6 terabytes onto a DVD-sized disk

FLU CONTROL
Talking tough saves lives

OBESITY Q&A
Trends and treatment

CROP IMPROVEMENT
Gene targeting advances

NATURE JOBS
Inspection syndrome

Abstractions

LAST AUTHOR

As every gardener knows, annual plants flower only once during their life cycle, and need to be replaced each year. Perennials, on the other hand, can flower many times, cycling between flowering and vegetative growth for many years, in time with the changing seasons. Although flowering in annuals has been extensively studied using the model plant *Arabidopsis thaliana* (thale cress), few studies have looked at the regulation of flowering in perennials. Maria Albani, a postdoc at the Max Planck Institute for Plant Breeding Research in Cologne, Germany, and her colleagues looked for genes that regulate perennial growth in *Arabis alpina* (alpine rock cress), a perennial relative of *Arabidopsis thaliana* that grows in mountainous regions. They identified a gene, *PEP1*, that controls flowering in response to low temperatures — ensuring flowering in the spring — and that also restricts the duration of the flowering season and enables the return to vegetative growth every year (see page 423). Albani tells *Nature* about the significance of this gene.

What drew you to this topic?

I'm interested in flowering in perennials — annuals die after flowering but perennials somehow find a way to trick the whole plant-senescence programme and keep growing vegetatively. This is one of the first studies to focus on the molecular mechanisms of flowering in perennials. We hope that *Arabis alpina* becomes the lab plant of choice for studying perennials, as *Arabidopsis thaliana* has for annuals.

Did you encounter any challenges along the way?

We began from zero. When we started, nothing was known about the way *Arabis alpina* grows, flowers or manages to be a perennial. On top of that, the plant's life cycle slowed down our experiments. It took at least four months from the time we planted the seeds for flowers to bloom. We had to do many experiments in parallel because it would have taken too long to wait for one experiment to end before starting the next.

Does your finding have practical applications?

The fact that *PEP1* affects the duration of a plant's flowering season is very important. Many fruit crops — for example, apples and grapes — are perennials. If we can understand how *Arabis alpina* cycles between flowering and vegetative growth every year, it will help us to understand how other perennials go through similar cycles. However, *PEP1* has mainly been found in species closely related to our model plant — not in fruits — and practical applications will require more extensive research. ■

MAKING THE PAPER

Joris De Ridder

Asteroseismology in red giants might yield clues to stellar evolution.

Probing the interiors of the Sun and other stars might sound tricky, but a technique similar to seismology provides a way. Just as seismic waves on Earth reveal information about the planet's interior, sound waves that travel through the Sun provide astronomers with clues about what is going on beneath its surface.

On page 398, Joris De Ridder, a postdoc at the Institute of Astronomy in Leuven, Belgium, and his collaborators reveal that 'asteroseismology' can be applied to the study of red giants — stars approaching the end of their evolution. "If you go to the doctor, he might listen to the sound of your heart to find out about its condition," says De Ridder. "In a similar way, we are listening to the sound of a star to understand its interior."

Sound waves travelling through a star can be detected as variations in the star's brightness. De Ridder and his co-workers had previously tried to detect such variations in red giants using Earth-based telescopes. But the results turned out to be the subject of some debate. One group of researchers, including De Ridder, suggested that the oscillations seen in these red giants were only radial — consisting of spherically symmetrical expansion and contraction — and would therefore contain little information. Others believed that some of the oscillations were non-radial — a type that is potentially a lot more useful for asteroseismology.

Because of this work, De Ridder was asked in 2004 to join the community of European and Brazilian scientists using France's Convection, Rotation and Planetary Transits (COROT) space telescope, launched in December 2006, and to head a group responsible for observing and analysing red giants. As soon as the data from COROT arrived, De Ridder and his colleagues realized that they contained evidence of non-radial oscillations. "I was very happy to be proved wrong," laughs De Ridder.

But the real work is just beginning. "We have collected the sounds giant stars make," says De Ridder. "Now we have to learn how to listen." To analyse the data, De Ridder and his team are collaborating with theoretical astrophysicists. The idea is to develop mathematical models that predict what stars' oscillations would look like on the basis of their internal properties, and then compare those models with observed data. "There was some theoretical work done before that suggested that non-radial oscillations would not be visible in red giants," says De Ridder. "So now we have to go back to the drawing board to figure out why we do see non-radial oscillations."

One detail that De Ridder would ultimately like to extract from the data is the density of a red giant's core, because that measurement relates to a star's evolution. "Red giants are elderly stars. Our own Sun will someday become a red giant. So we want to understand how this evolution occurs," he says.

De Ridder has now been asked to coordinate another project to observe red giants with NASA's Kepler telescope, which was launched into space earlier this year. The instrument will monitor the brightness of more than 100,000 stars for three and a half years. "Among other things, we hope to obtain information about how the stars rotate internally," says De Ridder. "Red giants rotate very slowly, so we need to observe them for a long time." ■

FROM THE BLOGOSPHERE

Only 600 or so hand-picked students get to attend the annual Nobel Laureate Meetings at Lindau in Germany, but anyone can 'virtually' attend a selection of historical lectures. Alison Abbott, *Nature*'s senior European correspondent, describes the history of these meetings on The Great Beyond (http://tinyurl.com/pczhrd).

Count Lennart Bernadotte, great-grandson of Sweden's King Oscar II, who awarded the first Nobels, launched the meetings in post-war Germany to encourage the country's isolated doctors and scientists. In 2005, the meetings were updated to allow students from around the globe to mingle with top scientists.

Now, 100 years after the count's birth, 11 lectures from historical meetings have been digitalized and made available through the meeting's website (www.lindau-nobel.de). "The cleaned up voice recordings, accompanied by an introduction and charming black-and-white photos taken in Lindau, bring legendary scientists to life," writes Abbott. Highlights include Rita Levi Montalcini speaking about human rights and the reclusive Paul Dirac's lecture on the gravitational constant. ■

Visit Nautilus for regular news relevant to *Nature* authors ▸ http://blogs.nature.com/nautilus and see Peer-to-Peer for news for peer reviewers and about peer review ▸ http://blogs.nature.com/peer-to-peer.

nature

www.nature.com/nature

Vol 459 | Issue no. 7245 | 21 May 2009

The female underclass

Funding agencies and universities should collaborate to make the most of women in research.

There's an old joke about the farmer who responds to a driver's request for directions: 'I wouldn't start from here'. For women beating a path to a leading position in science, 'here' would include Bulgaria, Croatia, Cyprus, the Czech Republic, Estonia, France, Greece, Hungary, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, the Slovak Republic, Slovenia and Turkey.

These countries are characterized in a report from the European Commission (EC) as being relatively inactive in national and institutional policies intended to overcome the obstacles faced by women in their scientific careers. More positively, the document details ways in which many of the main funding agencies across Europe are trying to improve matters.

The Gender Challenge in Research Funding (see http://tinyurl.com/ecgender) was written by a 17-strong expert group chaired by a woman and containing five men. That male minority is an inversion of the usual pervasive and regrettable imbalance of the sexes in European peer-review structures. Only in those countries that have been most proactive in supporting women's careers — Finland, Sweden and Norway — do women constitute more than 40% of 'gatekeeper' scientific boards, according to 2004 data, the latest available, quoted in the report.

Many leading funders are trying to do better. Germany's DFG, for example, has set equal opportunities as a statutory objective since 2002, with working groups targeting the various factors that undermine that goal. But Germany's overall performance is depressing for its women — and for its men too, who presumably want to see the country make good use of its talent. Between 1999 and 2004, the proportion of women acting as peer reviewers for the DFG rose — from 6% to 9%. Of all European Union countries, Germany has the lowest representation of women in the highest academic positions, despite an equal representation of men and women as graduates.

The pressures on women who want to excel in science are acute everywhere. This is particularly true for mothers of young children who, even in the most progressive countries, are generally expected to take on most of the responsibility for home and family while still being expected to write proposals, publish papers and spend long hours in the lab. Added to that is the committee work. Ironically, being a member of a minority that is targeted for positive action can lead to endless requests for advice and involvement, which cut even further into research time.

"The pressures on women who want to excel in science are acute everywhere."

Many of these pressures will ease only when fathers regard themselves as having equal responsibility for parenting. But employers also have a responsibility to assist parents. Another report published last week by the EC, *Women in Science and Technology — Creating Sustainable Careers* (http://tinyurl.com/womensci), highlights the ways in which Europe's employers provide support. These include such prosaic but essential initiatives as ensuring that important meetings are timed to allow parents to leave the office as necessary, and not overlooking those who work part-time when it comes to assigning senior responsibilities.

According to the report, the Netherlands is a notable hotspot for promoting women's interests. Over the past ten years, the funding agency NWO has given Dutch universities incentives to award senior lectureships and professorships to high-achieving women, without branding them as tokens.

Such collaboration, perhaps with sticks as well as carrots, between funding agencies and the institutions they fund, is essential if robust change is to come more rapidly. Without it, Europe will continue to include far too few countries that, for ambitious women scientists, are good places from which to start. ■

Can coal be clean?

New money must provide stimulus to get carbon capture and storage up and running.

There is good news about carbon capture and storage (CCS), the technology that is intended to slow global warming by capturing industrial carbon dioxide emissions and injecting them underground. Last week, US energy secretary Steven Chu outlined plans for using some $2.4 billion in economic stimulus money to research aspects of CCS. These efforts will join important research already under way. Several European nations are looking at technical issues in partnership with industry; Australia has a cutting-edge research programme; China has entered the game; and the United States has a number of existing pilot projects across the country.

Unfortunately, none of the current work translates into the rapid deployment required to prove this technology in the commercial arena. The different CCS technologies need to be demonstrated on power plants new and old, and industry must show that the CO_2, once injected into old oil and gas fields or saline aquifers, will stay put. Although Australia, China and Britain are working on such demonstrations, there is widespread agreement that many more are needed. Last year, the G8 leaders rightly called for upwards of 20 demonstration projects around the globe. Without that kind of commitment, no one will ever know what the true potential of CCS could be.

Chu seems to recognize the problem. The energy department is in talks to restart the United States' flagship CCS project FutureGen, a projected coal-fired power plant that would capture CO_2 and store it underground. The Bush administration shut down FutureGen last year after a dispute over rising costs, signalling to the rest of the world

that the United States might not be serious about this technology. But now that Chu has so much stimulus money for CCS at his disposal, FutureGen would be one good place to put it.

On Capitol Hill, the Senate Committee on Energy and Natural Resources also took a welcome step forwards last week by unveiling legislation that would remove regulatory hurdles for up to ten large-scale CCS projects. The bipartisan bill lays out a mechanism for the energy department to help shoulder the burden of risk and then eventually take over responsibility for long-term monitoring of the injected carbon.

The situation in Europe is similarly encouraging. The European Commission last year called for upwards of a dozen CCS demonstration projects, without explaining who would pay for them. That question was partly answered in December, when the European Union dedicated the revenues from 300 million of the allowances in its carbon-trading system. That could provide several billion euros at current carbon prices.

The United Kingdom took things a step further last month by announcing that it will require all new coal-fired power plants to have at least partial CCS. This sends a powerful signal that the days of unfettered coal burning are coming to an end. Other nations would do well to follow suit.

Some environmentalists have been hammering home the point that there is no such thing as clean coal, and they may be right. Even if operators did somehow manage to bury 90% of the CO_2 emissions from coal-fired power plants, that would still leave all the emissions and other environmental impacts from mining and transporting the coal itself.

Indeed, in an ideal world, burying CO_2 wouldn't be necessary. Civilization would instead rely on carbon-free energy resources such as solar, wind and nuclear power, and would reserve CO_2 for feeding algae and making carbonated beverages or, better yet, cement. But getting there will take time, and that is what carbon storage could provide. It's worth the effort. ■

Responsible interrogation

Psychologists have a moral duty to help prevent torture.

There are unequivocal points to be made within the debate now raging in the United States over the Bush administration's use of what it described in its sanitized parlance as 'enhanced interrogation techniques' to wring information from detainees suspected of terrorism — techniques better described as torture.

Despite plausible-sounding talk about 'states of induced dependency' and the like, there is no scientific basis for asserting that techniques such as waterboarding, or slamming people against a wall, are fast or effective ways of getting at the truth (see *Nature* **445,** 349; 2007). Indeed, it is hard to imagine any ethical way a controlled study on that question could be carried out. What is known to work — and surprisingly rapidly, according to field anthropologists, investigative journalists, police detectives and others with practical experience at getting information from reluctant or hostile sources — are the 'soft' methods of building rapport and trust.

And even if physical or mental torture could be shown to be effective in some immediate, tactical sense, that would be beside the point: torture is a violation of human rights and of international law, and is a threat to the long-term health of democracy. It is not to be tolerated.

Beyond that, there are few easy answers. Witness the struggles by the American Psychological Association (APA) to lay out ethical guidelines for psychologists involved in US national-security-related interrogations such as those that took place at Guantanamo Bay in Cuba.

The discussions, which were made public earlier this month on the non-profit news site ProPublica, were carried out on a confidential listserv in 2005 and involved the ten members of the APA Task Force on Psychological Ethics and National Security. Their work led to a set of 12 principles that were issued in a 2005 report (www.apa.org/releases/pens0705.html).

The most inflammatory issue, now that the task force's work has been thrust back into the limelight, is that six of its members were on the Pentagon's payroll. This might seem reasonable: guidelines should be informed by people who know what they're talking about. But it has led the Massachusetts-based activist group Physicians for Human Rights, among others, to charge the APA with having excessively cosy relations with the military on torture — or, at the very least, with letting the Pentagon dictate a set of guidelines to its own liking.

The evidence for this is not obvious in the 12 principles themselves. One forbids psychologists to engage in, direct, support, facilitate or offer training in torture or other cruel, inhuman or degrading treatment; another articulates a moral obligation for them to report acts of torture to the "appropriate authority". But the very fact that collusion charges have been made suggests how sensitive the subject is.

Another, long-standing issue for many APA members can be found in the first of the 12 principles, which explicitly states that it is ethical for psychologists to be involved in interrogations. Other professional societies have taken a less permissive tack; the American Medical Association, the American Psychiatric Association and the World Medical Association have all come out against having their members participate in interrogations.

But such restrictions fly in the face of the reality that interrogation is a necessity in preventing loss of life from terrorism, and that some professionals feel it is their duty to ensure that the activity is conducted responsibly. The risks of abuse are ever present, and having a professional present should serve as protection for detainees, provided the professional adheres to, and is held accountable to, the most fundamental medical ethic of all: 'do no harm'.

Mike Gelles, a task-force member who was at the time chief psychologist for the Naval Criminal Investigative Service, maintains that his active involvement at Guantanamo Bay allowed him to bring concerns about interrogation methods to military leaders there, leading them to change those methods. He deserves the last word. "Removing professional psychologists from these settings," he wrote in 2007 to colleagues who were calling for a moratorium on psychologists' involvement in interrogations, "will impact the degree of oversight and inevitably increase the likelihood of abuse, thus having precisely the opposite effect of what occurred as a result of my involvement at Guantanamo Bay." ■

RESEARCH HIGHLIGHTS

Middle Ordovician orgy

Geology **37,** 443–446 (2009)

Fossils of some of the largest trilobites ever found have been located in Portugal, and they bear witness to a very active social life in these extinct marine arthropods.

Artur Sá at the University of Trás-os-Montes and Alto Douro in Vila Real and his colleagues found fossils reflecting mass-mating, moulting and furtive manoeuvres among trilobites that lived during the Middle Ordovician period, 470 million–460 million years ago. The arthropod assemblage, from a roof-slate quarry in the Arouca Geopark, captures five contemporary families of three trilobite orders in a single formation for the first time.

Sá's team notes gigantism among six of these species, with one specimen reaching 70 centimetres long, and estimates for incomplete remains suggesting a possible 90 centimetres in another. The researchers suggest that their large size might be an adaptation to cold water.

M. VALÉRIO/AROUCA GEOPARK

GENETICS

Long-lasting without fasting

PLoS Genet. **5,** e1000467 (2009)

Restricting calories is known to increase life span, but changing the source of those calories can affect organisms in similar ways, reports a team led by Valter Longo at the University of Southern California in Los Angeles.

Longo's team studied long-lived *Saccharomyces cerevisiae* mutants lacking *SCH9* and other genes involved in yeast life extension. The mutants expressed higher levels of genes involved in glycerol synthesis and metabolism. Knocking out glycerol-synthesis genes eliminated life extension in mutants lacking *SCH9*. And a glycerol diet allowed yeast to live at least twice as long as those fed glucose, and slightly longer than those on calorie-restriction diets. The team concludes that replacing pro-ageing foods such as glucose with glycerol helps to boost cellular protection and lengthen yeast lifespan.

NANOMATERIALS

Inked in

J. Am. Chem. Soc. **131,** 6692–6694 (2009)

Like a pen writing with molecular ink, the copper-coated tip of an atomic force microscope can attach molecules from a solution to anchored molecules on a surface through chemical reactions, forming patterns with lines as narrow as 50 nanometres.

Fraser Stoddart and his colleagues at Northwestern University in Evanston, Illinois, achieved this precision by exploiting the bond-forming reaction between two chemical groups: azides and alkynes. Molecules decorated with these groups link in the presence of copper, so the tip acts as a moving, solid catalyst. It could be used to write arrays of modified biomolecules onto a surface.

A related 'dip-pen' technique coats the tip with a copper-containing solution, but has a lower writing resolution of 300 nanometres.

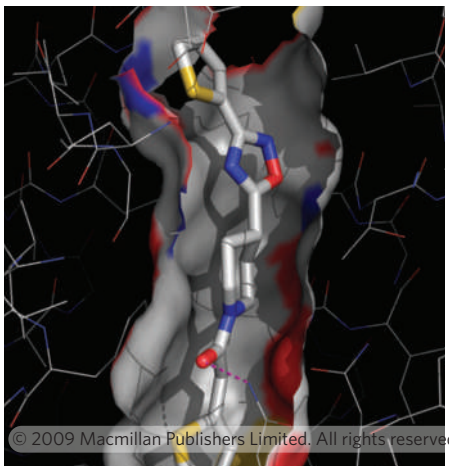
MICROBIOLOGY

Tag-teaming tuberculosis

Nature Med. **15,** 537–544 (2009)

High doses of the drug ethionamide cause side effects that limit its use against tuberculosis. Now, Alain Baulard of the Pasteur Institute in Lille, France, and his colleagues report that blocking the regulatory protein EthR in *Mycobacterium tuberculosis* boosts ethionamide's activity, markedly reducing the necessary dose of the drug.

EthR controls production of the bacterial enzyme EthA, which activates ethionamide. Baulard's team designed and screened a drug library for compounds likely to block EthR (one example pictured below) and thereby release its control on EthA. One such compound, BDM31343, bound to EthR and rendered ethionamide three times more potent against tuberculosis in mice.



CLIMATE

Cyclones take it higher

Geophys. Res. Lett. doi:10.1029/2009GL037396 (2009)

Tropical cyclones spew water vapour into the stratosphere, possibly adding to global warming, according to David Romps and Zhiming Kuang of Harvard University.

By analysing 23 years of satellite pictures and tropical-cyclone tracking data, they found that the storms frequently launch clouds and ice upwards. Whereas only 3% of non-cyclone cloud in the upper troposphere reaches the stratosphere, 8% of cyclone clouds reach those heights.

Stratospheric water vapour has an important role in global warming and ozone depletion. And because global warming may alter cyclone frequency and severity, this may represent another climate-change feedback mechanism.

CHEMISTRY

Mini magnets

J. Am. Chem. Soc. doi:10.1021/ja8098454 (2009)

Semiconductor nanoparticles, or quantum dots, should not be magnetic, yet some reports suggest that they are. Robert Meulenberg, currently at the University of Maine in Orono, Jonathan Lee at Lawrence Livermore National Laboratory in California and their co-workers present evidence that this magnetism comes from the chemical groups stuck to the dots' edges.

The group used X-ray magnetic circular dichroism and X-ray absorption spectroscopy to study the magnetic properties of electrons involved in chemical bonds in cadmium, one ingredient of cadmium–selenide quantum

dots. The authors say these observations reveal that cadmium–selenide quantum dots are paramagnetic — their magnetism is induced and temporary — owing to interactions between cadmium and the chemical groups added to its surface to stifle its reactivity. The team saw no evidence to substantiate previous claims that quantum dots are ferromagnetic — that is, permanent magnets.

EVOLUTION

Home-field advantage

Am. Nat. **173,** 579–588 (2009)

It has long been thought that for many species, local populations are adapted to their local areas and would show greater fitness there than anywhere else. However, some transplant studies have shown 'out of town' organisms doing better than the natives.

Joe Hereford, now at the University of Maryland, College Park, looked at 74 transplant studies containing 777 estimates of local adaptation. He found that species showed local adaptation 71% of the time, and that, in general, native populations were 45% fitter than transplanted populations. And species displaying high local adaptation were not always very unfit when transplanted.

NEUROGENETICS

Protecting plasticity

Nature Neurosci. doi:10.1038/nn.2327 (2009)

Angelman syndrome is a form of mental retardation caused by mutations in the gene *UBE3A*. Scientists have discovered a role for the Ube3A protein that might explain the learning deficits associated with the disorder.

Benjamin Philpot of the University of North Carolina at Chapel Hill, Michael Ehlers at Duke University in Durham, North Carolina, and their colleagues focused on the mouse visual cortex. They found that, unlike normal mice, when mice lacking Ube3A were exposed to light, they lost synaptic plasticity — the learning-associated ability to change the strength of signals sent between brain cells — in this area. But the Ube3A-deficient mice recovered plasticity when deprived of visual stimuli, suggesting that Ube3A may help to maintain experience-dependent plasticity during periods of high brain activity.

D. MANNING & B. GLOVER, UNIV. CAMBRIDGE

ECOLOGY

Pollinators get a grip

Curr. Biol. doi:10.1016/j.cub.2009.04.051 (2009)

Animal-pollinated plants produce many cues and structures that help pollinators to navigate plant parts. Most flowering plants have cone-shaped cells on the surface of their petals, but the specific function of these cells was unknown. An elegant experiment now shows that they help pollinators to get a grip on the plant surface.

Beverley Glover at the University of Cambridge, UK, and her colleagues observed the behaviour of bumblebees (*Bombus terrestris*) on natural and artificial surfaces that were coated with flat or conical cells. When the surfaces were presented at awkward angles, the bees grasped more easily and preferentially selected the textured surfaces. By making the flower surface tractable for landing pollinators, the authors suggest, conical epidermal cells increase the efficiency of pollination.

CANCER BIOLOGY

Cancer cop back on the beat

Cancer Cell **15,** 376–388 (2009);
Cancer Cell **15,** 441–453 (2009)

Reactivating the cancer-fighting protein p53 when it has been disabled in tumour cells can rein in cancer-promoting genes and kill cancer cells in culture.

The protein is inactivated in many human cancers, allowing tumour cells to escape programmed cell death. Galina Selivanova of the Karolinska Institute in Stockholm and her co-workers report that a compound called RITA suppresses the expression of several cancer-causing genes by reactivating p53. RITA also unleashes a host of cell-death-promoting proteins, killing cancer cells.

Meanwhile, Klas Wiman, also of the Stockholm Karolinska Institute, and his colleagues show that another p53-reactivating compound, PRIMA-1, converted to reactive compounds in living cells. One of the active metabolites reacts with the sulphur-containing groups on the p53 protein, possibly restoring mutated p53 to a normal conformation. p53 is then able to trigger death in cancer cells.

JOURNAL CLUB

William C. Hwang
Burnham Institute for Medical Research, La Jolla, California

A structural biologist has great expectations for llamas' small antibodies.

Llamas aren't just unusual and exotic looking: their antibodies are also a reason for much excitement. Made entirely of heavy chains, they are about half the size of those found in humans and many other vertebrates, which are normally composed of both heavy and light chains. When it comes to therapeutic applications, these larger antibodies are hard to store and deliver. But llama and other camelid antibodies demonstrate superior heat-stability and solubility, without compromising affinity or specificity, making them an attractive alternative.

Robin Weiss of University College London and his colleagues isolated three llama antibodies, known as 'neutralizing' antibodies, that can broadly prevent multiple HIV subtypes from infecting cells (A. Forsman *et al. J. Virol.* **82,** 12069–12081; 2008). They began by creating an antibody library from two llamas immunized with the HIV gp120 antigen. To select for neutralizing antibodies, antibodies were raised against one HIV subtype but cross-screened against multiple subtypes. The researchers also included a competitive elution step to select antibodies that can compete with binding by CD4, the primary HIV receptor on human T cells. It remains to be seen how these neutralizing antibodies fare in animal studies and where they bind in atomic detail.

Intriguingly, there have been reports of several potent, broadly neutralizing human antibodies (for example, F10 and CR6261 against influenza's haemagglutinin) in which only heavy chains are involved in antigen binding — reminiscent of the situation of llama antibodies. These studies corroborate that the heavy chain alone can mediate broad neutralizing activity, and invite speculation that this may be a special strategy engaged by the human immune system to reach cryptic binding sites. Llama antibodies may be even better suited for those hard-to-reach targets.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

NEWS

J. MEAD/SPL

Solar sell: research into renewables gets limited funding in Australia's budget.

Sunny outlook for Australian science

Research programmes win big in budget, but critics say environment is 'overlooked'.

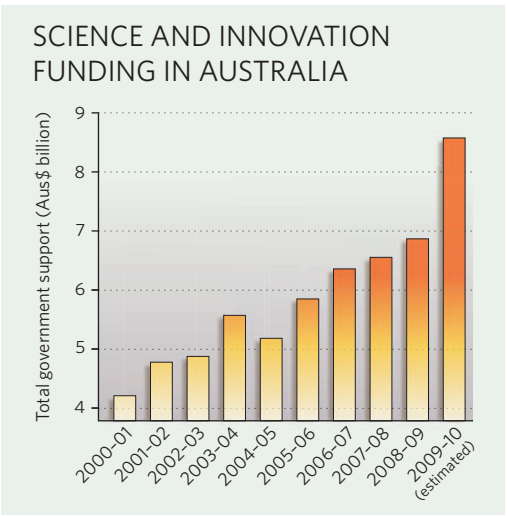
SYDNEY

An unexpected funding boost has left many Australian scientists in a positive frame of mind after federal treasurer Wayne Swan delivered a budget last week that raised research and education expenditure by almost 25% over the previous year's spending (see graph).

The budget, presented to parliament on 12 May, features a raft of research-related spending measures, including funding reform for universities and an overhaul of tax arrangements for private-sector research and development (R&D) that had been recommended by recent government reviews of innovation and higher education. The budget numbers still need to be passed by the Senate.

Altogether, the proposed funding measures are worth Aus$5.7 billion (US$4.3 billion) over four years, including Aus$3.1 billion on R&D, with the remainder on education. "In the lead-up to the budget, the government talked down expectations," says John Quiggin, an economist at the University of Queensland in St Lucia. "But the budget outcome is far better than might have been feared in the light of pre-budget softening up."

The 25% increase from 2008–09 to 2009–10 would represent the highest annual rise since records of science funding began in the 1970s, says Ken Baldwin, president of the Federation of Australian Scientific and Technological Societies, the main representative body for Australia's scientists. "I think this puts us back in the running with other countries that have invested significantly in science," he told *Nature*.



SOURCE: AUSTRALIAN GOVERNMENT

The investment is particularly significant given the current global financial downturn, Baldwin says. "Australia is relatively well placed in an economic sense, and is now relatively well placed in a research and development sense," he says. "It shows that the overall commitment to use research and development as a driver for economic recovery is front and centre of the government's planning."

A key measure is a phased commitment to fund the full cost of research in universities. Researchers have long complained about not getting enough funding to cover indirect research costs.

Carbon concerns

Not everyone was completely happy with the budget details. Some environmentalists, including prominent scientist and author Tim Flannery, have recently criticized Prime Minister Kevin Rudd's centre-left Labor government for failing to deliver on environmental promises, such as how a cap-and-trade programme to reduce carbon emissions might be implemented.

The budget does set aside Aus$4.5 billion for clean-energy initiatives, including Aus$1.6 billion for solar power. Still, Aus$2.4 billion is earmarked for low-emissions coal technologies.

"What we got from the budget were some nice token gestures, but no policies that would allow a roll-out of renewable energy on a large scale," says Mark Diesendorf, an environmental scientist from the University of New South Wales in Sydney. "The problem is that the coal industry and some of the other big greenhouse polluters have enormous political influence in Australia," he says. "The federal government has actually rejected implementing policies that would drive large-scale solar and other large-scale renewable-energy technologies that have enormous potential."

COMFORTING CASH Handling money soothes physical pain and the hurt of rejection. www.nature.com/news

PUNCHSTOCK

Yet for most researchers, the increased spending was welcome. For the CSIRO, Australia's national science agency, the budget offered an increase in direct funding of Aus$43 million, or 6.3%, in the next financial year. The agency will also gain Aus$120 million to deliver a new marine-research vessel by 2012, and Aus$30 million to expand its Living Atlas of Australia, an integrated system for managing the country's biodiversity data.

CSIRO chief executive Megan Clark says that, among other things, the budget increases will boost research into agricultural productivity.

In astronomy, the budget promises Aus$80 million to support Australia's bid to host the Square Kilometre Array, a Aus$3-billion radio-telescope project for which it is competing with South Africa. It also includes Aus$20.9 million to help Australia take sole responsibility for the Anglo-Australian Observatory when the United Kingdom withdraws its involvement next year, and Aus$40 million for an Australian space-research programme.

The private sector also gained from the budget, with biotechnology firms winning an item that was top of their wish list: reforms to taxation of R&D. Michelle Gallaher, chief executive of the biotech industry group BioMelbourne Network, says that many in the sector would see the benefits of the announced tax credits, worth an estimated Aus$1.4 billion per year. "The budget certainly was better than our expectations," she says.

In medical research, cancer specialists were particularly pleased by news that the government may spend Aus$560 million to build a network of regional cancer centres. Patients treated in those centres would be well situated to participate in clinical trials, says Ian Olver, chief executive of the Cancer Council Australia.

Despite the positive mood, some researchers, such as Robert Graham, president of the Association of Australian Medical Research Institutes, are already warning that big gains may not be sustainable in future years.

"Now that we're finally in the running in the biomedical science and biotechnology industry," he says, "we look to the Rudd government to ensure that in the future we are not relegated back to a high chair." ■

Stephen Pincock

CDC

H1N1 update

As confirmed by the World Health Organization, as of 19 May:
Cases: 8,829, in 40 countries
Deaths: 74, in 4 countries
For more on the H1N1 flu, see pages 322–325 and online at www.nature.com/swineflu

Public donations to lift research

Researchers looking for a little extra money to explore an idea may soon have a new funding 'agency' to ask for help: the masses. A Florida researcher has launched a project designed both to solicit public donations for individual research programmes and to inspire public interest in science and engineering.

The 'SciFlies' project (www.sciflies.org) will profile scientists from a range of disciplines and the new ideas they want to pursue, or ways in which they would like to expand their current research programme. Website visitors will be able to donate any amount to support the projects they find most interesting or worthwhile.

David Fries, a marine engineer at the University of South Florida in St Petersburg, conceived of and heads the SciFlies effort. His main inspiration was long-standing frustration with a research funding structure that, with few exceptions, offers scientists no intermediate steps on the way to requesting full grant funding. So last year he joined forces with Larry Biddle, a political and non-profit fund-raiser who helped to pioneer Internet-donations campaigns for Howard Dean's 2004 presidential run, and Michelle Bauer, founder and chief strategist of Common Language, a communications firm based in St Petersburg.

D. FRIES

David Fries: hoping his concept of grass-roots funding will fly.

Pretty fly for a sci guy

Fries hopes that SciFlies will be a way for scientists around the world to secure funding in the US$5,000–10,000 range. Such funding can be enough, for instance, to hire a graduate student to run initial experiments that might determine if an idea is worthy of a grant proposal.

Fries says the name SciFlies has a triple meaning. It's a homage to the ubiquitous fruitfly research model, a shorthand description of the goal to create a 'swarm' of science supporters, and a reference to 'fly', a slang term for cool.

The site will profile research projects and the scientists themselves, including their overall research vision, past accomplishments, interests and even what they are currently reading. The hope is to inspire genuine and long-term interest in researchers, who will be required to post regular updates on their work, and science in general.

For now, the team plans to profile every scientist who signs on. Anyone who doesn't bring in donations or keep their profile updated is likely to be culled from the list.

Oscar Schofield, an oceanographer at Rutgers University in New Brunswick, New Jersey, will be one of the first scientists to be profiled. By using tools such as robotic gliders, his group gathers data, including ocean temperature and salinity, that are used regularly by the public. But in the past he has had to turn down offers by users to donate to his programme because of the difficulty of processing small donations. "What's brilliant about SciFlies is not only its ability to reach people, but how it will provide a mechanism where those little contributions build up and then get delivered at some later point," he says.

SciFlies will distribute cheques to researchers after their donations reach $5,000. Anyone who doesn't make that amount will not get any money, and donors will be asked to redesignate their money.

Rush Holt (Democrat, New Jersey), a physicist and member of the US House of Representatives, sees the SciFlies concept as akin to public funding of the arts, and a potential way to boost research in line with increases proposed by the administration of US President Barack Obama. "It's an innovative approach that I think we should be encouraging," says Holt.

An initial SciFlies website, aimed at getting researchers to sign on for profiling, launched on 15 May. The full site, ready for donations, will be launched in July. ■

Mark Schrope

Research from rubble

University returns to work in makeshift lecture halls and laboratories.

The historic centre of L'Aquila remains off-limits. G. MELINO

L'AQUILA, ITALY

Amid the rubble of the University of L'Aquila, Italy, which was mostly destroyed by a magnitude-6.3 earthquake on 6 April, a Nobel-prizewinning biologist arrived last week to offer his support. To a packed marquee, Robert Horvitz tried to restore a semblance of normalcy by lecturing about his work on programmed cell death.

"It was a question of showing scientific solidarity with people who live in a community that has suffered losses on so many levels," says Horvitz, of the Massachusetts Institute of Technology in Cambridge. "They've lost life, homes — and also a perspective on how they can return to normality."

Fifty-five students were among the 295 people who died in the quake, which shattered the town and its neighbouring villages, and destroyed much of the university.

Six weeks later, with 70% of its staff homeless, the 23,000-student university is starting to work again — in tents or in buildings loaned by other towns. The underground particle-physics laboratory at Gran Sasso, which remained undamaged 15 kilometres from L'Aquila, resumed work on 4 May, even though 90% of its staff are homeless.

Researchers are worried about being forgotten by the outside world, or abandoned by their students, during the rebuilding process — which is expected to take many years. To help to avert a drain of human capital, the Italian government has promised to maintain a constant budget for the university, at €68.5 million (US$92 million) annually for three years, and to reduce student fees. It is also giving €70 million for rebuilding. Prime minister Silvio Berlusconi has also transferred July's G8 meeting from the island of La Maddalena to L'Aquila, to bring in external money for building, although some fear that it could prove to be a disruption.

To try to maintain meaningful international connections, the science faculty — about 465 of 600 academics — is offering the town of L'Aquila as an 'open laboratory' for the testing of new scientific ideas during reconstruction. "The university needs to be involved in the rebuilding of an ancient city with modern technologies and modern ideas," says Paola Inverardi, dean of science, who is living in a tent in her sister's garden. "The tragedy can also be an opportunity."

More than 1,300 scientists have indicated their interest in being involved on the 'Ideas for L'Aquila' webpage (www.ideasforlaquila.org). Ideas being discussed include a consortium for 'adaptive music', which would incorporate music that adjusts to a changing environment in the infrastructure of the restored city.

Inverardi, a computer scientist, says that collaborations forged with international partners could help the university to emerge from the disaster in better intellectual shape than before.

But the road to recovery is very steep. Only two buildings on the university's two out-of-town campuses remain structurally sound and will be habitable within a few months. The rest are substantially damaged — as is the entirety of the humanities faculty and university administration located in the historic centre. The centre is still a no-go area, a dusty, rubble-strewn ghost town brightened only by red fire engines. The yellow stuccoed university rectory seems to lean, crumbling into a narrow alley already choked with fallen bricks.

Each department has had to find its own interim solution for teaching, which began again a couple of weeks ago. Some classes are taking place in the sea of blue tents erected by the civil-protection service on campus, or in larger marquees that have also hosted exams and academic ceremonies — as well as the seminars of Horvitz and his wife, neurobiologist Martha Constantine-Paton.

The physics faculty found a relatively easy solution by moving into the above-ground facilities of the Gran Sasso laboratories, where many homeless staff also sleep. "Of course there will be crowding — and it will be for some years," says Gran Sasso director Eugenio Coccia. "But we are glad to be able to have such a role."

It has not been easy to find the mental energy to think about science in the circumstances, admits Gran Sasso physicist Francesco Arneodo. "With so many homeless it is hard to focus your full attention on research," he says,

Radon 'prediction' of earthquake rattles local scientists.

"There are always people who say they had predicted every earthquake," says Tom Jordan, director of the Southern California Earthquake Center in Los Angeles. But the city of L'Aquila, Italy, had a particularly troublesome claimant this spring in amateur seismologist Giampaolo Giuliani.

After an earthquake devastated the town on 6 April, the international media covered Giuliani's lament that Italian authorities had ignored his prediction of a big quake centred on Sulmona, a town some 60 kilometres away, at the end of March. Giuliani, a technician, had been tracking radon emission from the ground with a device he had patented.

Jordan has been invited by the Italian government to head an international committee of experts to analyse local seismic data — and, in part, to put to rest public concern that lives could have been saved by heeding predictions. Radon is indeed emitted from Earth during seismic activity, he says, but it is not useful in predicting earthquakes.

Giuliani's claim offended professional geologists such as Gianluca Ferrini of the University of L'Aquila, who had been monitoring the mounting seismic activity for several months. "We knew that the epicentre of what we were recording was getting deeper and stronger, so we knew there was a possibility of a big earthquake," he says. Ferrini and his colleagues had been giving survival advice in schools. If people had been evacuated from Sulmona to L'Aquila in response to an inaccurate prediction, he points out, many more people would have died. **A.A.**

D. HERSCHKOWITZ

DEFENDING ISRAELI SCIENCE
Minister speaks out on proposed budget cuts.
www.nature.com/news

the strain clear on his face. "But now it is OK — we are back!"

Geologist Gianluca Ferrini finds his attention still absorbed in his work as a voluntary member of the science faculty's civil-protection unit. He was one of the first on campus after the earthquake hit at 3:32 a.m., checking for immediate dangers of fire or flooding from gas leaks or broken water pipes. He remembers that even in those eerie night hours, looters were already trying to make off with computers.

He also recalls, with some pain, how in the dawn light the unit quickly moved on towards town to help the bare-handed search in the rubble for survivors, or for bodies. As the morning wore on, Ferrini says he found six of the dead, their dust-covered arms indistinguishable from the grey rubble except by touch.

Ferrini also helped to secure many of the university's scientific resources — from keeping fridges running after electricity was lost, to looking after experimental animals and stowing away valuables such as a collection of insects that had once belonged to Charles Darwin. "I have very little time to do my research work," he says, although he does frequently drive more than two hours to different locations to teach.

For Horvitz, visiting L'Aquila was a deeply moving experience. "Life does go forward — and these scientists are trying to find out how to deal with their lives and at the same time continue research," he says. ■

Alison Abbott

Elements reveal fossils' origins

Researchers are closing in on a way to 'fingerprint' fossils to match them to their precise geological origins.

Federal land managers and scientists have long dreamed of finding a way to link a potentially stolen fossil to a specific site, and thus subvert poachers who take fossils from public land. The US Paleontological Resources Preservation Act, signed into law in March, contains tough criminal penalties — felony charges carrying two- or five-year sentences — for such offences.

At the eighth Conference on Fossil Resources on 20–21 May in St George, Utah, researchers were scheduled to report on methods that can closely tie a fossil to a site through analysis of rare earth elements (REEs).

There are 15 REEs in the lanthanide series — from lanthanum, the lightest, to lutetium, the heaviest — all of which can be incorporated into bones as they become fossilized. The ratio of REEs varies from place to place depending on the chemical composition of groundwater in the region; a fossil's REE signature can thus be matched up to a geological location like pieces fitting into a jigsaw puzzle.

"Rare-earth-element analysis could be used as a prosecutorial tool."

The methods are getting ever more precise, says Dennis Terry, a geologist at Temple University in Philadelphia, Pennsylvania. "I have been able to say if a fossil is from Nebraska or adjacent South Dakota with 99% confidence," he says. "But now we are getting much closer to distinguishing locations on a regional level."

In a study to be presented in Utah by Terry's undergraduate student William Lukens, the team matched 45 fossil samples to three sites from which they were taken with 96.7% accuracy. All of the fossils originated just a few kilometres apart in and around Toadstool Geologic Park, a remote area in northwestern Nebraska where fossils are often poached.

Terry has also been using REE analysis to date bones more precisely, by comparing their elemental ratios to the stratigraphy of the rocks from which the specimens were taken (D. E. Grandstaff and D. O. Terry *Appl. Geochem.* 24, 733–745; 2009). This summer, he and his colleagues will be testing this method on the Flagstaff Rim, south of Casper, Wyoming, where they hope to use REE analysis to reduce the dating error bars from a span of about 100,000 years to about 20,000 years.

Meanwhile, Celina Suarez, a doctoral student at the University of Kansas in Lawrence, was due to present results at the conference showing how a laser can chart REE signatures in different parts of a single fossil specimen, making identification more exact. Rather than dissolving a piece of fossil in solution and testing the subsequent samples for REEs, her technique involves burning numerous spots on a specimen — with each ignition emitting gas that a mass spectrometer reads for REEs. This method can produce an elemental map of a bone's REE absorption.

"There are some problems, but overall the method is sound," says Suarez. "In some cases, I can tell where a fossil was removed from within a quarry."

Lucy Kuizon, head of palaeontology for the US Bureau of Land Management in Washington DC, says that such approaches are one way to fight fossil poaching. "These techniques are still under study, and not perfected yet," she says.

However, REE analysis is already moving into the courts. In a fossil probe on 11 April, the bureau used the technique to support a search-warrant application that was approved by a federal court in Kansas City, Missouri.

One day, REE analysis could be used as a prosecutorial tool, says palaeontologist Barbara Beasley of the Nebraska National Forest in Chadron. ■

Rex Dalton

D. TERRY

Fossils incorporate chemical elements from soil.

SPECIAL REPORT

The planetary police

Planetary scientists are looking for new ways to sterilize their spacecraft, so that they won't be excluded from exploring interesting places. **Eric Hand** reports.

In 1975, the twin Viking landers sped off to Mars as the most sparklingly clean things ever put into space. If either was to have any chance of detecting microbial life in its scoops of Martian soil, it couldn't risk carrying stowaways from its launchpad at Cape Canaveral. "Going to Mars to study Florida is not a very good idea," says John Rummel, NASA's former planetary protection officer.

All the cleaning ended up being overkill. Viking found Mars to be cold, dry and dead. The rules for planetary protection — the endeavour of keeping Earth bugs off Mars, and Mars bugs off Earth — were relaxed. The Mars landers that followed were built in clean rooms and wiped down with alcohol, but did not have to be fully baked and sterilized.

But now engineers are looking at ways to make the rigour of Viking-type sterilization de rigueur. This year, scientists at the Jet Propulsion Laboratory in Pasadena, California — where many NASA spacecraft are built — will submit documentation to NASA arguing for the adoption of a second method of sterilization, which uses lower-temperature vaporization of hydrogen peroxide rather than high-temperature baking. If adopted, the protocol could make it easier for engineers to design and manage fully sterilized spacecraft. "We're trying to make it easier to do upfront," says NASA's acting planetary protection officer Catharine Conley.

Tough microbes

Sterilization is crucial, because in recent decades ever more hardy microbes have been discovered on Earth (see 'Extremophiles on Earth'), even as water ice has been found in ever more places on Mars. That raises the potential for similarly tough microbes to live in places on Mars, probably underground, that may yet be wet. Last week, the US National Research Council released a report reaffirming that, if and when researchers manage to bring back a rock sample from Mars, it will need to be kept in a state-of-the-art, maximum-security biocontainment laboratory — just to avoid the possibility of anything escaping.

Planetary protection efforts have mostly focused on the possibility of Earth microbes forward-contaminating other Solar System bodies. International guidelines for sterilizing spacecraft are set by the Committee on Space Research (COSPAR), which divides missions into five broad categories ranging from flybys of bodies not of interest to astrobiologists, to a sample return from another world. Within these categories — and subdivisions for various targets of interest, such as Mars, or Jupiter's moon Europa with its ice-covered ocean — are strict rules for the number of spores that may be cultivated off the spacecraft after it has been heat-shocked at 80 °C for 15 minutes.

For the most part, scientists haven't minded such restrictions, because extra-clean instruments make for more precise measurements. But engineers baulk at the cost and complexity of full sterilization; baking both Vikings, for instance, cost 10% of the US$800-million mission, or about $320 million in today's dollars. And so they prioritize, choosing to clean just the instrument that will potentially come in contact with the extraterrestrial surface.

> **"If you're going in with a blunt tool, and you really need a scalpel, there's no point in going in there until you get the right equipment."**

For example, Phoenix, the NASA spacecraft that landed last summer near the planet's northern polar plains, had its sterilized robotic arm and scoop entombed in a 'biobag' during transit, so as to minimize the chance of passing on Earth microbes when it scraped through a thin layer of soil to find ice. But the lander set down in millimetres of ice exposed by retrorockets during landing. Now, some claim[1], Phoenix's legs may have become covered in droplets of briny water created when deliquescent salt lying above the subsurface ice was kicked up during landing. Any microbes that might have clung to the lander could thus find a relatively comfortable home.

"I would have wanted the lander legs to be sterilized," says Conley. Phoenix had qualified as being clean, though, because only its arm needed sterilizing.

Full sterilization through heat baking is the rule for any spacecraft destined for 'special regions' on Mars — areas where ice or water could be near the surface, with temperatures warm enough for organisms from Earth to replicate. A 2006 National Research Council report argued that, given the potential for transient water to be in many places on Mars, the entire planet should be treated as 'special' until scientists can show otherwise. That suggestion, however, didn't go down very well. "Certainly, the community didn't want all of Mars to be considered a special region," says Conley.

Later that year, Mars researchers, using models and data from orbiters, did their best to identify the likely zones of near-surface ice and gullies that could indicate transient water in the near past. These are the designated 'special regions'. The regions between 30° north and 30° south qualified as not-so-special[2] (see map). But since then, there have been mid-latitude discoveries of buried glaciers[3] and tiny pools of ice excavated by recent asteroid impacts[4] — suggesting that the non-special zone might be even smaller than currently designated.

The $2.3-billion Mars Science Laboratory (MSL) super-rover, due to launch in 2011, has a radioactive heat source that could melt ice

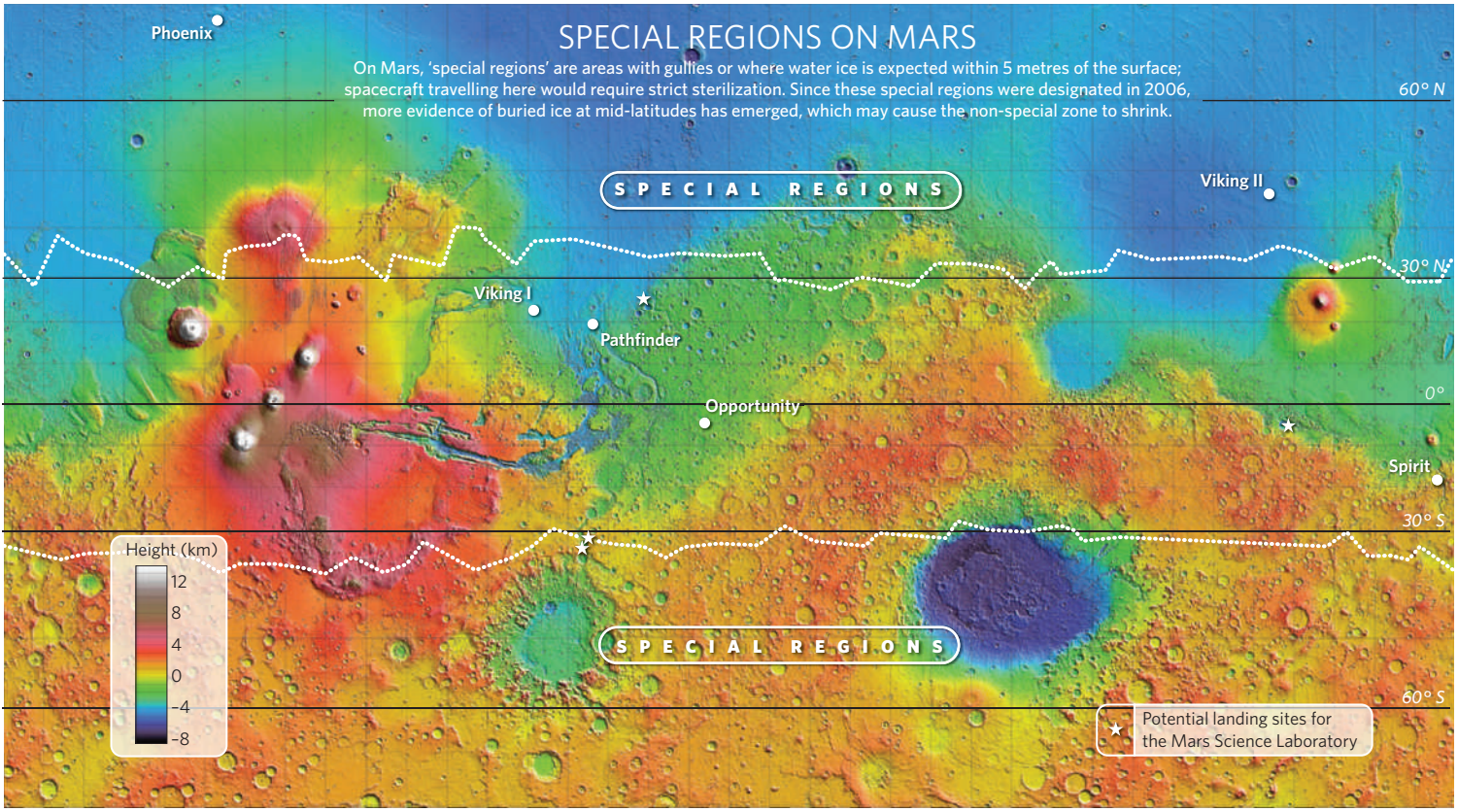
EXTREMOPHILES ON EARTH

Life has been found everywhere, from boiling hot springs to Himalayan glaciers to nuclear reactors. These extremophiles on Earth make life on Mars more of a possibility — and also make planetary protection more of a challenge.

Parameter	Example
High temperature (thermophile)	Archaeal strain 121 (121 °C)
Low temperature (psychrophile)	Himalayan midge (–18 °C)
High pH (alkaliphile)	*Alkaliphilus transvaalensis* (pH 12.5)
Low pH (acidophile)	*Ferroplasma acidarmanus* (pH 0)
Salt-tolerant (halophile)	Halobacteriaceae (saturated solutions 10 times saltier than sea water)
Lack of oxygen (anaerobe)	Methanogens
Desiccation (xerophile)	Lichens, cyanobacteria
Radiation (radiophile)	*Deinococcus radiodurans* (withstands radiation 1,000 times the lethal dose for a human)
Pressure (piezophile)	Obligate bacterial strain MT41 (1,000 times atmospheric pressure)
Vacuum-tolerant	Tardigrades ('water bears'), insects, seeds
Longevity	Bacteria isolated from 250-million-year old salt crystal

J. L. AMOS

CAMBRIAN CHEMISTRY
Animals on the sea floor may have stoked global sulphate levels.
www.nature.com/news



for many years — complicating things if there were a crash or mishap above thinly covered ice. Mission managers elected to avoid special regions altogether, figuring that there would be less chance of contamination if the ice were deep. Although the MSL is capable of detecting organic molecules such as methane, it is not technically considered a life-detection mission — because that, too, would bump it up to a level that requires full-baking sterilization for those instruments. (One of the reasons that mission managers for Phoenix and the MSL have said their probes are exploring 'habitability' rather than searching for life, Conley says, is to avoid stiffer requirements.)

Engineers steering the probe's descent have other reasons to avoid higher latitudes, and the shortlist of the four places where the MSL may end up are all in equatorial regions. But it is possible that the rover could stumble on a surprise — a hydrothermal vent, say, or a pond of buried ice not too far below the surface. In one of these scenarios, the rules could force scientists to throw the rover into reverse. Scientists would be told to pull out, "even though it was scientifically meaningful, and even though it would be desirable", says Karen Buxbaum, planetary protection manager for the Mars programme at the Jet Propulsion Lab.

Would scientists have the patience to leave the area in a pristine state, being forced to wait years for another, cleaner probe to come along? "I don't think people would have a big problem with it," says John Mustard, a geologist at Brown University in Providence, Rhode Island, and chair of a NASA Mars exploration advisory committee. "If you're going in with a blunt tool, and you really need a scalpel, there's no point in going in there until you get the right equipment."

Getting cleaner

ExoMars, a Mars astrobiology rover proposed for launch in 2016 by the European Space Agency (ESA), will be the cleanest Martian mission since Viking. But last year, ESA decided that ExoMars would also avoid special regions, so that only the life-detection instruments rather than the whole spacecraft would need to be heat-baked, according to ESA planetary protection officer Gerhard Kminek. Areas containing features such as gullies, although interesting, would have been difficult to land near anyway, he says. "It saves money, but it also was a practical issue."

Rummel says that scientists have already shown restraint about exploring sensitive regions. In 2003, mission controllers sent Galileo, a mission to the Jupiter system, plummeting early into Jupiter rather than risk losing control of it and having it hit Europa and its ice-capped ocean. "We knew that to protect Europa we had to kill the spacecraft," says Rummel, now director of the Institute for Coastal Science and Policy at East Carolina University in Greenville, North Carolina.

For now, Conley and her colleagues are working to ensure that, whatever craft leaves this planet, they know what is on it. As part of modernizing and updating the planetary protection techniques, she hopes to replace the decades-old diagnostic for measuring spacecraft contamination, which relies on counting the numbers of bacteria as a proxy for overall microbial life. By adding approaches involving biochemistry and the polymerase chain reaction, she says, scientists can generate a more specific inventory of exactly what types of microbes, if any, are present. Recent studies along these lines have shown how several species of archaeal bacteria still cling to surfaces in spacecraft-assembly clean rooms[5].

Even with existing technology, she notes, scientific exploration and planetary protection should not be in conflict. "There is nothing that is off limits," she says. "You just have to be clean enough." ■

1. Renno, N. O. *et al. 40th Lunar Planet. Sci. Conf.* www.lpi.usra.edu/meetings/lpsc2009/pdf/1440.pdf (2009).
2. Beaty, D. *et al. Astrobiology* **6,** 677–732 (2006).
3. Holt, J. W. *et al. Science* **322,** 1235–1238 (2008).
4. Byrne, S. *et al. 40th Lunar Planet. Sci. Conf.* www.lpi.usra.edu/meetings/lpsc2009/pdf/1831.pdf (2009).
5. Moissel, C. *et al. ISME J.* **2,** 115–119 (2008).

Alzheimer's theory makes a splash

Neuroscientists probe idea that neuronal pruning may contribute to degenerative disorder.

The Alzheimer's research community is buzzing about a theory suggesting that a close relative of the β-amyloid protein, and not necessarily β-amyloid itself — the long-standing suspect — may be a major culprit in the disease.

The theory holds that an amyloid-related mechanism that prunes neuronal connections in the brain in the fast-growth phase of early life may be triggered by ageing-related processes in later life to cause the neuronal withering of Alzheimer's disease (A. Nikolaev *et al. Nature* **457,** 981–989; 2009).

"We have yet to get a disease-modifying drug that works. So we're missing something, and maybe this is one of the missing pieces," says Donna Wilcock, a neurologist at Duke University in Durham, North Carolina.

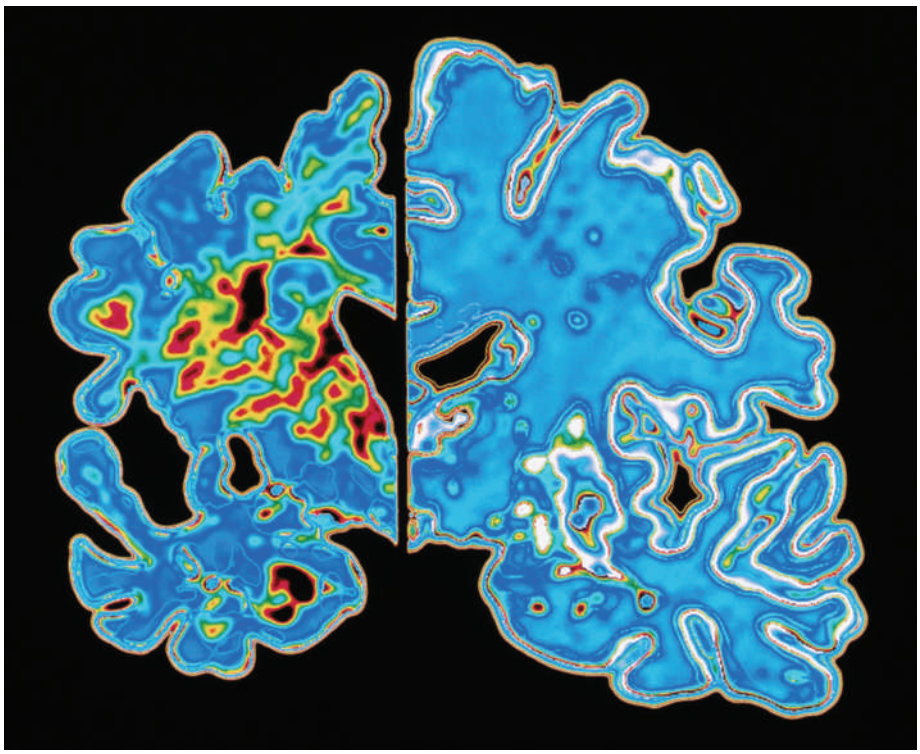
"I think people are bored of the amyloid hypothesis and would just love to have something else to follow up," adds John Hardy, a neurologist at University College London.

Insoluble clumps of the β-amyloid peptide appear in the brains of patients with Alzheimer's, and mutations in amyloid's precursor protein (APP) have been linked to rare, familial forms of the disease. But how amyloid contributes to the damage of Alzheimer's is not clear, and several anti-amyloid drugs have failed in phase III clinical trials (see *Nature* **456,** 161–164; 2008). Some scientists are unsure that any form of β-amyloid contributes to much of the neuronal destruction.

Now, researchers suggest that Alzheimer's might be caused by an APP-derived protein — just not necessarily β-amyloid. The paper describes a naturally occurring brain-pruning process and identifies one of its key players: N-APP, a fragment of APP that is adjacent to β-amyloid and is cleaved from APP by one of the same enzymes.

N-APP triggers the self-destruct pathway by binding to a neuronal receptor called DR6 (death receptor 6). DR6 is highly expressed in the human brain regions most affected by Alzheimer's, so it is possible that the N-APP/DR6 pathway might be hijacked in the ageing brain to cause damage.

The lead author of the study, Marc Tessier-Lavigne of Genentech, a biotechnology company in South San Francisco, California, plays down the conflict with the amyloid hypothesis. He emphasizes that even if N-APP does contribute to the withering of neuronal connections in Alzheimer's, β-amyloid might play a part too, perhaps by depressing synaptic function, as other research suggests. "The two are not mutually exclusive," he says.



A. PASIEKA/SPL

The brain of a patient with Alzheimer's disease (left) is shrunken compared with a normal brain (right).

In any case, N-APP's relevance to Alzheimer's is not yet proved. Bruce Yankner, a neurologist at Harvard University, says that the study focused on embryonic cell types, which are not affected in Alzheimer's disease. (Tessier-Lavigne says that his lab hopes to publish data soon on N-APP/DR6 mechanisms in Alzheimer's-relevant adult brain cells.)

New model army needed

There is also no evidence, so far, that the N-APP pruning pathway is active in animal models of Alzheimer's disease. The most commonly used mutant 'Alzheimer's mice' overexpress APP and mimic the amyloid deposition seen in Alzheimer's brains, and therefore presumably overexpress N-APP, too. Yet these mice strikingly don't show Alzheimer's-like neuronal destruction.

"It opens up a lot of lines of new research for us."

The usefulness of these mice as models for human disease has long been questioned, points out Tessier-Lavigne. In initial studies, N-APP triggered self-destruction only when neurons were also deprived of the nerve growth factors that normally keep them healthy. Such deprivation might not occur as much in mouse brains as in aged human brains, he says; the team is now studying mice that under-express such growth factors.

Other research groups are studying the N-APP pathway for links to Alzheimer's. David Holtzman, a neuroscientist at Washington University in St Louis, Missouri, is testing blood and spinal fluid from patients with Alzheimer's to see if "this fragment of APP is even present in human body fluids". If it is, Holtzman hopes to find out whether it can be used in diagnosing Alzheimer's or tracking its progress.

Wilcock says that her research group recently developed transgenic mice that have Alzheimer's-like amyloid deposits, neurofibrillary tangles and neuronal loss, and she hopes to study whether the N-APP pathway is driving some of that destruction. "I think it opens up a lot of lines of new research for us," she says. "These are important questions that everybody should be looking at." ■

Jim Schnabel

Canadian charged with smuggling Ebola

The arrest at a US border of a researcher allegedly trying to smuggle non-infectious Ebola DNA in from Canada is raising questions about high-containment lab security.

But officials on both sides of the border are saying little about how the samples were removed from the Canadian National Microbiology Laboratory in Winnipeg, or about the suspect, 42-year-old Konan Michel Yao, who was a fellow at the laboratory until January.

On 5 May, Yao was discovered to have 22 vials of Ebola genetic vectors as he tried to drive into North Dakota. He was charged three days later with smuggling biological material and making false statements to authorities. The incident became public only last week.

Yao, now a Canadian citizen but born in Ivory Coast, told US agents that he was heading to a fellowship at the National Institutes of Health (NIH) in Bethesda, Maryland, and had documents indicating this to be the case, court records say.

An NIH spokeswoman refused to discuss the case. Canadian government officials didn't respond to an inquiry.

New York's health commissioner to head CDC

C. HONDROS

Thomas Frieden.

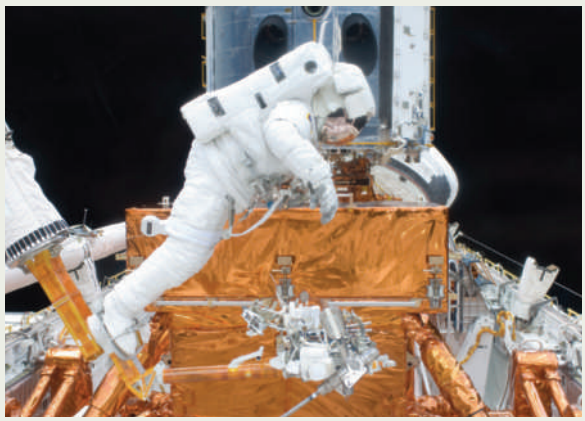
Thomas Frieden, New York City's health commissioner, has been appointed head of the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia.

Frieden worked for the agency from 1990 to 2002, when his work included tackling the spread of multidrug-resistant tuberculosis in India. In New York, he has worked to reduce tobacco use and to develop emergency-response plans in case of a pandemic. Back at the CDC, he will take over an agency on the front lines of battling the H1N1 swine-associated influenza outbreak.

During the 1990s, Frieden worked for another New York health commissioner: Margaret Hamburg, who on 18 May received Senate confirmation to head the US Food and Drug Administration.

Frieden's appointment does not require Senate confirmation.

Astronauts complete Hubble repairs



NASA

The Hubble Space Telescope has had its vision refreshed. On 18 May, astronauts completed the final overhaul of the ageing astronomical icon. The mission has included installing a new camera, a new spectrograph and making repairs to two other instruments (see *Nature* **459,** 21; 2009).

The servicing — involving five days of spacewalks (pictured) from the *Atlantis* space shuttle — also included new batteries, gyroscopes and repairs to the telescope's thermal insulation. It was the fifth servicing mission, the first in seven years, and also the last, expected to keep Hubble working until at least 2014.

But another NASA observatory, the Spitzer Space Telescope, ran out of coolant on 15 May. The infrared telescope will only be able to perform limited tasks when its instruments are calibrated for warmer operation. It happened just one day after the European Space Agency successfully launched an infrared telescope, Herschel, that will build on Spitzer's discoveries. Planck, a second satellite launched with Herschel, will study the cosmic microwave background.

Lawsuit targets validity of human-gene patents

In a case that could determine the fate of gene patents, a group of cancer patients, clinicians, researchers and activists have sued the US Patent and Trademark Office and the owners of patents on two genes associated with cancer.

The lawsuit, filed on 12 May, targets Myriad Genetics and the University of Utah Research Foundation, both based in Salt Lake City, and both of which own patents on the *BRCA1* and *BRCA2* genes. Certain mutations in these genes have been linked to an increased risk of breast and ovarian cancer but, because Myriad Genetics holds exclusive rights, diagnostic testing for these mutations can only be performed at Myriad labs.

The lawsuit asserts that patents on human genes associated with breast and ovarian cancer are invalid and that they violate the US constitution. The lawsuit "does have a lot of potential", says Arti Rai, a law professor at Duke University in Durham, North Carolina. "But it's still a hard case to make."

Re-election of Indian prime minister aids nuclear deal

A nuclear trade deal struck last year between India and the United States remains on course to be implemented, after the ruling left-of-centre United Progressive Alliance (UPA) secured an unexpectedly decisive victory in India's elections.

Opposition parties had vowed to review the deal if they came to power, arguing it would reduce the independence of India's domestic nuclear programme.

Aside from nuclear policy, the contesting parties broadly agreed on scientific issues. But the continuity assured by the election of prime minister Manmohan Singh to a second term in office means that the science agenda can move forwards without disruption, says Rajagopala Chidambaram, principal science adviser to the government.

The UPA has committed to raise government spending on science from 0.9% of gross domestic product to 2% by 2012.

For a longer version of this story, see http://tinyurl.com/q6cggs.

Austrian scientists celebrate CERN U-turn

Austria has reversed its decision to withdraw from CERN, Europe's particle-physics laboratory near Geneva in Switzerland.

On 8 May, science minister Johannes Hahn had announced that Austria would end its participation in 2010. Instead, the country would use its annual contribution of about €17 million (US$23 million) to make up shortfalls in the country's research budget. The decision sent shock waves through the country's physics community (see *Nature* 459, 151; 2009).

Protests were not in vain. On 18 May, Chancellor Werner Faymann overruled his science minister and announced that Austria would continue its 50-year-old membership in CERN without cuts.

"Austrian science as a whole will benefit from this firm declaration of belief in basic research," says Christian Fabjan, director of the Austrian Academy of Sciences' Institute for High Energy Physics in Vienna.

TIME TO CHANGE THE BULB

The incandescent light bulb is being phased out, but what will replace it? **Stefano Tonzani** investigates the technologies that are vying for our sockets.

The Centennial Light, which hangs in a fire station in Livermore, California, is the oldest working light bulb on Earth. The four-watt night-light was switched on in 1901 and has been shining almost non-stop ever since, consuming roughly 3,500 kilowatt-hours of energy in total. As the picture opposite shows, the bulb also looks surprisingly familiar: the technology of incandescent lights has changed very little over its lifetime. Inside the bulb is a filament — carbon in this case, tungsten in today's models — that is heated by the flow of electricity until it glows white and lights up the room. The design is simple, versatile and cheap, just as it was when Thomas Edison first made it a commercial success in the 1880s.

Nonetheless, that technology is now on the way out. In today's energy-hungry world, the devices are too wasteful: some 98% of the energy input ends up as heat instead of light. Halogen lamps, which look more high-tech, are not any better. Multiply that waste by the number of incandescent bulbs in residential, industrial and commercial settings — an estimated 4 billion standard light sockets in the United States alone — and it is clear why several countries are seeking to eliminate the bulbs entirely as a way to control carbon dioxide emissions. In 2007, for example, Australia became the first country to ban incandescent bulbs entirely; the phase-out is scheduled to be completed by 2012. The member states of the European Union agreed to a similar ban in 2008. And the United States has pledged to eliminate most incandescents by 2014.

"The lighting field is a fairly conservative one, so these government mandates are putting on some welcome pressure to evolve," says Karl Leo, an optoelectronics specialist and a founder of Novaled, a company in Dresden, Germany, that develops organic light-emitting diodes (OLEDs).

> **"Government mandates are putting welcome pressure on the field to evolve."**
> — Karl Leo

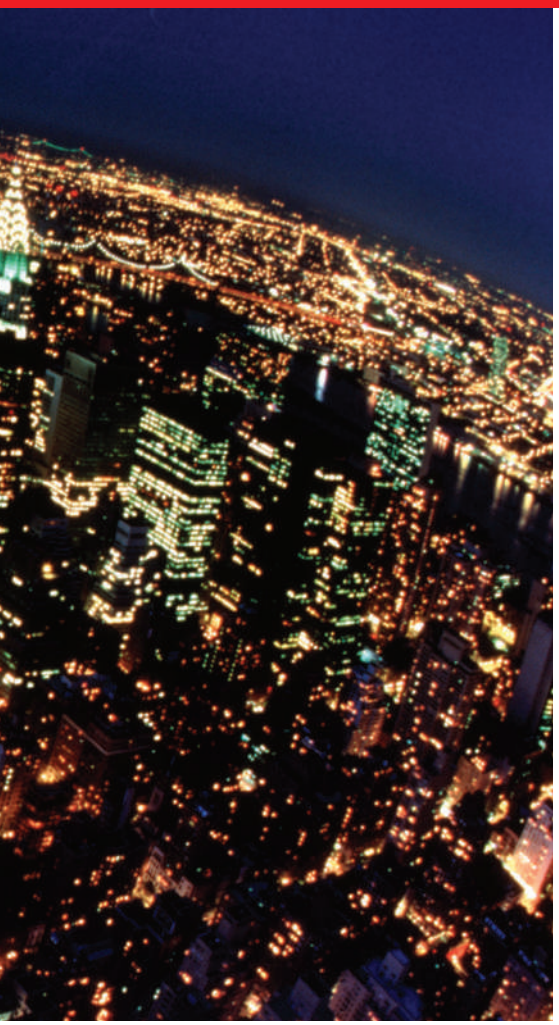
But evolve into what? Although getting rid of incandescent bulbs makes environmental and economic sense, the race for a long-term replacement is wide open.

At present, the only technology that is mature enough to take over from the conventional light bulb is fluorescent lighting, which can turn 10–15% of the input energy into light. Fluorescent technology has improved substantially since the days when it was synonymous with being harsh and funny-coloured, and has come to dominate in industrial and commercial settings, where energy efficiency and long life are prime concerns. In recent years, compact fluorescent bulbs that can be screwed into standard sockets have brought it ever farther into the home.

But fluorescent lighting has a number of drawbacks. For example, fluorescent lamps do not work well in cold temperatures, and their lifespan can be significantly shortened if they are turned on and off frequently. Perhaps worst of all, each lamp contains a small amount of mercury, which is toxic. This presents consumers with a disposal problem at the end of the lamp's life.

Some people also still complain about the colour rendering of fluorescent lights — the way the lamps make objects look compared with their appearance in natural sunlight. Despite the substantial progress, domestic users in particular tend to prefer the warmer, slightly red-tinged tones of incandescent lights — although that preference is highly individual, says Charles Hunt, an electrical engineer

OWAKI-KULLA/CORBIS



at the University of California, Davis. "The subjective fondness for particular light shades depends on gender (women tend to prefer less harsh, warmer-coloured lights) and origin of the person (people from northern European countries prefer warmer lights whereas southern Europeans prefer colder, more blue-tinged ones)." Another issue is that fluorescent lights require special circuitry to work with a dimming switch. And dimmable is desirable: "Fifty per cent of home lights in the United States are dimmable," says Hunt.

These problems may be overcome — but they seem serious enough to encourage innovators to search out successor technology.

Heavy investment

Perhaps the most widely anticipated of the technologies vying for centre stage is the light-emitting diode (LED), which consists of two types of semiconductors in contact. When a voltage is applied, positive charges coming from one side flow towards the junction and meet negative charges coming from the other side. As these charges combine, they release their energy in the form of light, usually as one particular colour.

LEDs are long-lived, robust and roughly twice as efficient as fluorescents. Indeed, they are already widely used for computers, television sets and other consumer electronics, and are becoming a market leader for outdoor applications such as traffic lights and indicator lights on cars. "There are so many advantages to LEDs that we think there lies the future of lighting," says Hans van Sprang, a senior scientist at Philips Research Laboratories in Eindhoven, the Netherlands. Philips and other big industry players are investing heavily in the technology, supporting materials-science research that has helped LED technology to evolve rapidly.

D. JONES

Going strong 108 years on.

Despite this, LEDs have not yet been adopted on a large scale for general lighting applications. One problem is that an LED light powerful enough for room lighting has a very high initial cost compared with an equivalent incandescent bulb. This can be a large psychological barrier for consumers, even though the cost of energy and maintenance is considerably lower. Another problem is that an LED's lifetime can reduce dramatically if it is operated at a high temperature. This makes heat dissipation an important issue, especially for powerful lamps, and complicates efforts to reduce costs. Making the LED semiconductors from substrates other than sapphire would be cheaper, and the alternatives, including silicon-based substrates, might improve heat management.

Another challenge is how to generate white light from LEDs. The preferred technique for commercially available devices is to coat a blue or ultraviolet LED with a phosphorescent material that will absorb the monochromatic emissions, and then re-emit the energy as a broad-spectrum white light. Another, potentially more energy-efficient, way of generating white light is to mix the light of red, blue and green LEDs. But both of these methods have issues with colour rendering. And the latter method has the added problem that the lifespans of the three different LED types are not the same, so the light will change in colour as the lamp ages. Gain in one dimension and you can lose in another — devices that have very good colour rendering tend to have poor energy efficiency.

One potential solution is now under development by Sandra Rosenthal, a chemist at Vanderbilt University in Nashville, Tennessee. Her idea is to use an ultraviolet-emitting LED to energize the electrons in cadmium selenide nanocrystals, which respond by re-emitting a white light with very good colour rendering. "This could be a viable alternative if we could substantially improve their efficiency," says Rosenthal.

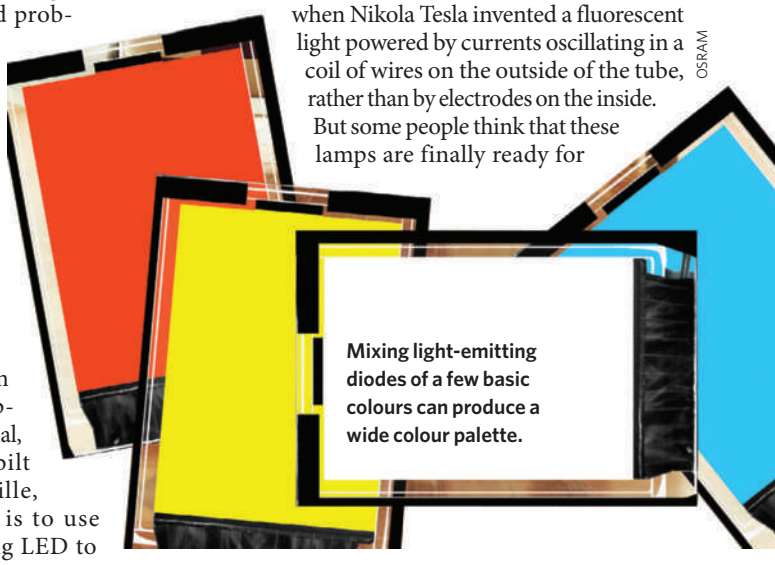
Organic compounds, which have already been looked at as a possible alternative to silicon in solar cells, are being investigated for use in LEDs. OLEDs produce light in much the same way that ordinary LEDs do, except that the positive and negative charges originate in organic compounds rather than in crystalline semiconductors. Typically, these organic compounds are attached to a fixed polymer sheet. The advantage of organic materials is that, at least in theory, they can be produced comparatively cheaply with the same roll-to-roll technology used to handle other types of plastic films.

The main problem with OLEDs is that the organic materials are degradable by water and oxygen, which tends to give the devices a short lifespan. This can be solved, to some extent, by encapsulating the organic compounds in an inert, transparent polymer such as epoxy resin. But the compounds degrade intrinsically anyway, especially the blue OLEDs that are required for mixing with red and green to generate white light.

The outsiders

Farther out of the mainstream — in the sense that the technologies are being developed by smaller start-up companies — are induction lamps and cathodoluminescence.

Induction lamps, also known as electrodeless lamps, have been around since the 1890s, when Nikola Tesla invented a fluorescent light powered by currents oscillating in a coil of wires on the outside of the tube, rather than by electrodes on the inside. But some people think that these lamps are finally ready for



OSRAM

Mixing light-emitting diodes of a few basic colours can produce a wide colour palette.

J. GOLLINGS/ARCAID/CORBIS

Light-emitting diodes, such as these in Craigieburn, Australia, can be combined to create huge displays.

prime time. The newest devices feature an electrodeless bulb that is filled with argon gas plus a small amount of metal halide salts. A microwave generator, much like the ones in microwave ovens, produces a wave that is channelled through a waveguide and concentrated on to the container, where it ionizes the gas to form a plasma and vaporizes the salts. The plasma and vapour together generate a broad-spectrum white light with an efficiency similar to that of LEDs. The devices are also very bright, which means they are likely to find their initial applications where intense light is required, such as in car headlights or industrial illumination.

"It will take LEDs a long, long time to catch up with the intensity of illumination possible with induction lamps," says Robin Devonshire, chief scientist of Ceravision, a company in Milton Keynes, UK, that is one of several developing this technology. As these lamps do not place electrodes in contact with the harsh plasma environment, they can potentially last for decades.

Cathodoluminescence works like the cathode-ray tubes found in old-fashioned television sets. It uses a source of electrons to bombard a phosphorescent material coated on the inside of a glass bulb, causing the material to emit light. An electrical field, high temperature or photoelectric effect is used to make a metal surface emit the electrons. Such a light source can be quite efficient, comparable to compact fluorescent sources. It renders colours well, and the lamps can be shaped to look like incandescent bulbs. Initial applications will be geared towards home usage.

"It will take LEDs a long time to catch up with the light intensity of induction lamps."
— Robin Devonshire

However, both induction lamps and cathodoluminescence lamps have a perception problem. "There is scepticism in industry with respect to these technologies that are not solid state," says Bruce Pelton, director of engineering of the University of California, Davis, California Lighting Technology Center. This is partly because solid-state devices such as LEDs — which generate light through processes in solid material, having no moving parts or bulbs that can break — are thought to have major advantages in the long run when it comes to ruggedness and long life. But it is also because a previous incarnation of the induction lamp, based on sulphur, failed to gain a foothold in the market. The sulphur lamp was efficient and bright, but was large and required air cooling to stop parts melting in the high temperatures reached by the sulphur plasma.

Blazing competition

The competition to replace incandescent light bulbs is likely to be fierce. But consumer acceptance is far from guaranteed for any of the rival technologies. One problem is the confusion generated by the sheer number of alternatives. Another is that each of these devices has several parts, so that the lifespans and energy efficiencies reported for basic technology do not correspond to those of the whole device, which are, as yet, not that far ahead of incandescent bulbs. In LED lamps, for example, the electronics or the phosphors are likely to degrade much earlier than the solid-state device itself. US Department of Energy data published in 2008 found that commercially available LEDs were about half as efficient as compact fluorescent lights. And, although development since then has been fast, the problems remain.

Meanwhile, because widespread acceptance of these technologies is crucial to changing the habits of consumers — and ultimately, to saving substantial amounts of energy — governments are keen to avoid the errors made with previous technologies, such as the early fluorescent lamps, which many end-users hated. If a technology is initially viewed negatively by the public, this can mar its subsequent evolution — a good reason not to mandate a technology before it is market-tested and ready. Comments on websites mentioning the planned phase-outs of incandescent lights have highlighted that many people's opinions of fluorescent lights have not changed much over time. Cost, colour rendering, flicker and the presence of mercury are only a few of the issues mentioned. Fluorescent lights are still a small percentage of the market compared with more energy-intensive lamps.

For all these reasons, the general-purpose incandescent light bulb might not be replaced by a single new source, but by a range of technologies, each suited to a particular use. For example, if OLED lighting can economically be produced in continuous sheets by industrial roll-to-roll techniques, it will be a natural candidate for flat panels that generate a diffuse glow for area lighting. That would make OLEDs a natural complement to the bright, directional light coming from semiconductor LEDs, which could instead be used for more light-intensive tasks such as reading. Such combinations could lead to new concepts of lighting design, so that architects could help save energy by not wasting light where it is not needed. ■

Stefano Tonzani is an associate editor at *Nature*.

NATURE|Vol 459|21 May 2009

PAINTING BY D. PARKINS

Nascence man

Like an alchemist of yore, Mike Russell is taking basic elements and trying to transform them — not into gold, but into the stirrings of life, **John Whitfield** reports.

You could call the two linked aluminium containers in Mike Russell's lab the biological equivalent of a particle accelerator. But rather than simulating the birth of the Universe, he hopes that this apparatus will recreate the first moments of life on Earth, and give experimental support to his ideas about how geology begat biology. Alternatively, you could call it a machine for making 4-billion-year-old waste.

One of the containers holds a liquid that mimics the oceans of the early Earth. The water is rich in carbon dioxide and iron, has a pH of 5.5 and is held at room temperature. The other container is heated to 130 °C, and its water is laden with hydrogen and sulphide. With a pH of 11, this second fluid is meant to stand in for the hot waters that spewed out of ocean-bottom springs early in the planet's history. The liquids mix in a chrome steel pressure barrel containing a catalyst of iron and nickel sulphide.

It is here that Russell (illustrated above) hopes to reproduce life's first steps, by reacting the carbon dioxide in the 'ocean' water with the hydrogen in the 'spring' water to make the simple organic molecules methane and acetate. Step by step, he thinks, the chemistry of life accreted around this reaction, until eventually, like caravels from the court of Henry the Navigator, the first cells carried it around the world.

As a candidate for the spark of life, this reaction has a lot going for it. It releases chemical energy and can fix carbon — that is, convert carbon dioxide into organic compounds — two of life's most characteristic properties. It uses ingredients that are abundant and it fits with what we know about the early Earth. Moreover, it is still used today, albeit with a good deal more sophistication, by the microbes called methanogens and acetogens, which produce methane and acetate as waste.

Russell has spent nearly three decades developing this hypothesis. Now, working at the Jet Propulsion Laboratory (JPL) in Pasadena, California, he's gearing up to test it, hoping to score a victory for the school of origin-of-life researchers known by the label 'metabolism first' in its long struggle with the more popular school called 'replicator first'. The latter holds that life began with a molecule — perhaps RNA or a simpler precursor — that was able to duplicate itself. Russell, however, thinks that the key breakthrough was the set of metabolic reactions that underpins biochemistry. The thermodynamic and chemical properties of the early Earth, he says, made these reactions a statistical inevitability.

Despite the rivalry, however, there is widespread recognition that Russell brings a much-needed dose of geological reality to research into the origin of life. "What I respect most about Mike's work is his keen insight into the early Earth's geochemical environ-

ment," says Robert Hazen, a geochemist and origin-of-life researcher at the Carnegie Institution for Science in Washington DC. "The origin of life is the story of the emergence of complexity, and you can't have the emergence of complexity unless you have a complex environment. Mike, as much as anybody out there, has recognized this fact and incorporated it into his models. That's his strongest contribution."

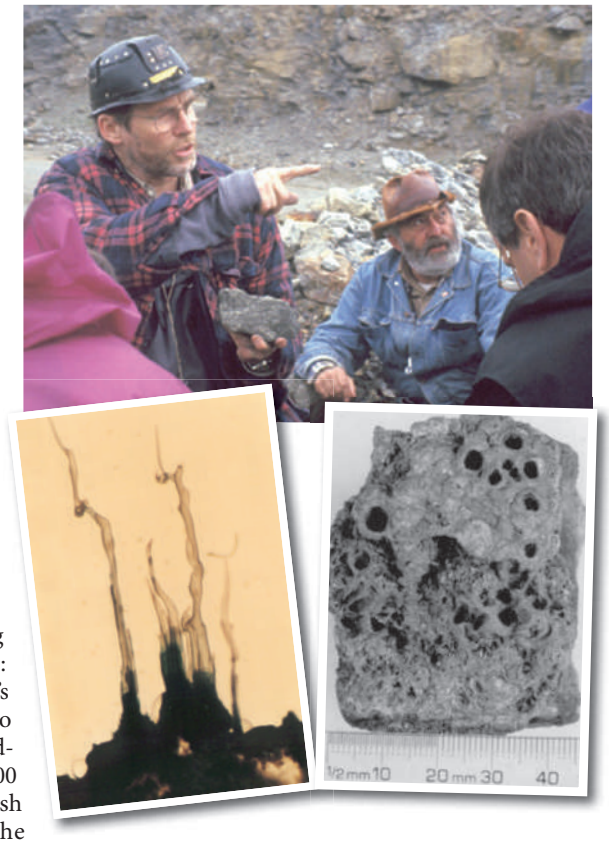
From aspirin to volcanoes

Studying the origin of life is not a great way to build a scientific career, so the people who work on the topic tend to have built up their reputations in other disciplines. But even by these standards, Russell's route to the JPL has been circuitous. When he left high school in 1958, he got a job making aspirin at a chemical plant in Ilford, a small town on London's northeast edge. But he continued to study at evening classes, and the Quaker-run company gave him days off to attend college.

Five years later, having left the factory and taken a degree in geology and chemistry, he found himself on the Solomon Islands in the Pacific Ocean, as a volunteer geologist with the UK Mission to the United Nations. In his first week, his boss pointed out the office window to a smouldering volcano on a nearby island. "He said: 'The chief has just radioed. He thinks it's going to explode.' I had no idea what to do," Russell recalls. Charged with deciding whether to evacuate the island's 3,000 inhabitants, Russell gave himself a crash course in volcanology, measured the ground temperature around the island's smoking volcanic vents, and decided, correctly as it turned out, that on this occasion the chief was mistaken.

While he was on the Solomons, Russell worked with the Australian geologist Richard Stanton from the University of New England in New South Wales. On Stanton's advice, he became an ore geologist, going to Canada to work in mineral exploration, before moving into academia in the late 1960s. Stanton also recruited Russell to his then-unorthodox idea that mineral deposits were the legacy of ancient submarine hot springs. When such hydrothermal vents were found on the Pacific floor in 1977, Stanton was proved right. Many valuable mineral deposits are indeed the remains of ancient vent sites, showing various similarities to the 'black smokers' that pour out water heated to 400 °C and are loaded with dissolved zinc, copper, iron and other elements.

By then, Russell was working at the University of Strathclyde, UK, and doing fieldwork in the Republic of Ireland at the mineral deposits in Silvermines, County Tipperary. There, he and his students found rocks riddled with small tubes of iron sulphide. They looked like miniature versions of the hydrothermal chimneys formed by minerals precipitating out of vent water.



G. EARLS

M. RUSSELL

J. H. E. CARTWRIGHT *ET AL.* J. COLLOID INTERFACE SCI. 256, 351–359 (2002)

Mineral deposits Mike Russell (in a helmet) found at Silvermines in Ireland (above, right) looked like chimneys near ocean hot springs. Chemical gardens (above, left) helped to inform his ideas.

A child's discovery

Russell began trying to work out what sort of environment would give birth to these structures. His suggestion that the tubes were formed in vents[1] met with a cool reception, he says, because the chimneys at black smokers were much bigger — 10 centimetres across, whereas those he saw in Ireland were less than 1 millimetre. Revelation came courtesy of Russell's 11-year-old son, Andrew. Russell had introduced Andrew to chemical gardens, toys in which pretty structures grow from a seed crystal added to a mineral solution. In a fit of destructiveness, however, Andrew had locked himself in the bathroom and started pulling the gardens apart. "Suddenly he yelled out 'Hey Dad, these things are hollow'," says Russell.

"I realized our little chimneys at Silvermines were actually chemical gardens," he says. That, he decided, must mean that black smokers were not the only kind of vent. There had to be cooler, gentler springs that would produce more delicate structures. More-or-less simultaneous with this thought came the idea that such a place was a good candidate for life's nursery. Some researchers had already suggested that hydrothermal vents had provided life's primordial source of energy and chemicals, but others protested that the extreme heat of a black smoker would break any large organic molecule into pieces. At the kind of vent Russell had in mind, however, the temperature would not have got much above 100 °C, which is much more amenable to organic chemistry.

The final clue emerged during a visit to Yugoslavia in the mid 1980s, which revealed even greater diversity in ancient hot springs. The water in modern black smokers is acidic, thanks to dissolved sulphur compounds creating sulphuric acid. When life emerged in Hadean times some 4 billion years ago, the ocean, too, would have been acidic, thanks to the large amounts of carbon dioxide in the atmosphere that would have dissolved in the waters. But in the Dinaric Alps, Russell saw deposits of the magnesium-bearing mineral magnesite that had formed in the ancient ocean floor and precipitated out of alkaline springs[2]. At the time, however, no active alkaline vents were known, only black smokers.

Between the late 1980s and the mid 90s, Russell and his colleagues stitched all their evidence into a portrait of the 'Goldilocks' spot, where the mineral chemistry was just right for what happens in cells today. "We recognized that some types of mineralization had a lot in common with the chemical processes of life," says geologist Allan Hall of the University of Glasgow, UK, and one of Russell's principal collaborators at the time.

Their theory of how life got going starts inside the tiny mineral chimneys. This protected environment would allow chemicals to become concentrated — a key problem facing anyone trying to explain how biochemistry can begin without cells. When these chimneys formed, they would have been gels rather than rocks, with membranes that would have allowed small molecules to pass through, much as cell membranes do. And the team found they could produce such a mineral gel in the lab[3]. The gel's membranes contained mineral sulphides of iron and nickel that would have catalysed organic reactions — just as these

metal sulphides do inside modern enzymes.

Across the membranes, gradients would have developed. Inside the vent, the water would have been hot, alkaline and rich in hydrogen, thanks to reactions between water and iron minerals in the crust, a process called serpentinization. Outside in the ocean, the water would have been cold and acidic. The vast majority of modern cells power much of their chemistry by creating similar gradients across their membranes. But they use a large variety of proteins to do it. Life's diverse ways to harness a ubiquitous energy source — the proton gradient — makes Russell think that the proteins are a secondary adaptation, and that life latched onto inorganic proton gradients before it could make its own.

Left to their own devices, hydrogen and carbon dioxide form methane only slowly, because the reaction's initial steps, from carbon dioxide to formaldehyde, require an input of energy. In the ancient vents, the energy of the proton gradient accelerated this reaction, says Russell. He draws an analogy with geological convection, the churning currents of ductile rock that speed up the release of heat from Earth's interior to the surface. "Metabolism is to geochemistry as convection is to geophysics," he says.

Russell initially thought that the key reaction in the origin of life was a redox reaction involving iron, so called because the iron is said to be reduced and the hydrogen oxidized. In 1998, however, he saw a paper in *Nature* arguing that eukaryotes arose when a hydrogen-requiring archaeal cell engulfed a hydrogen-producing bacterium[4]. Piqued by the mutual interest in hydrogen redox reactions as a biological energy source, Russell explained his ideas to one of the authors, William Martin. Martin, now at Heinrich Heine University in Dusseldorf, Germany, loved Russell's ideas. "I looked at it," says Martin, "and I said 'Well, this is easy — the origin of life is basically solved.'"

Martin had one problem. He thought that, if you want to say anything about how life came to be, then modern organisms must harbour the descendent of the first biochemical reaction, and in modern organisms the key to success is reducing carbon dioxide, not iron. "Is it reasonable to assume that what was possible for the very first cell has since been forgotten, and nobody is left who can do it?"

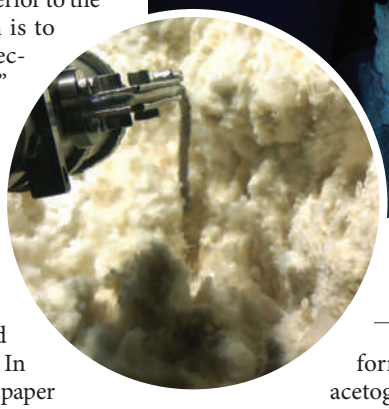
He steered Russell away from pursuing a hypothetical, forgotten reaction and towards what's called the Wood–Ljungdahl pathway, also known as the acetyl coenzyme A (CoA) pathway after the energy-rich molecule that forms its end product in modern acetogens and methanogens. Martin favours this explanation because it is the only one of the five known biochemical pathways for fixing carbon that turns a net profit in ATP, the universal chemical fuel of life. In this scenario, relatively cool spots within the vent favour the formation of acetate, an intermediate in the pathway, whereas in hotter spots, the reaction goes all the way to methane — possibly creating the divide between acetogens and methanogens, and between the groups that eventually became the Bacteria and Archaea.

Another advantage of the Wood–Ljungdahl pathway is that its products and intermediates plug into other metabolic pathways. Some of those make amino acids and nucleic acids, through reactions with ammonia — which Russell is including in his simulated vent water — and phosphate, which is present in his simulated ocean. When nucleic acids and amino acids first formed on Earth, their initial job, say Russell and Martin, would have been to catalyse reactions involving carbon dioxide and hydrogen[5].

In a metabolism-first world, before genetic molecules had a decisive role in evolution, selection would have favoured not the best replicator, but the reaction that sucked in fuel the quickest, denying energy to other chemical processes. And what would become the network of cellular chemistry could have grown by adding links that increased energy consumption. Even though there were no organisms, a set of reactions would store information in its components and processes. And that network could be said to replicate by drawing in more molecules and more energy into itself, a process sometimes called chemical evolution.

What's needed at this point is some evidence that will help to distinguish between the various hypotheses, says Robert Shapiro, a chemist at New York University. "Basically, one has to provide a set-up and demonstrate self-sustaining and evolving chemical cycles." Thanks to NASA's astrobiology programme, which has provided increased visibility to origin-of-life research, Russell is now in a position to attempt that as a researcher at the agency's JPL.

UNIV. WASHINGTON



NOAA

The carbonate structures at the Lost City hydrothermal field in the Atlantic Ocean have a delicate structure when first formed (inset).

Giant chemical garden

Russell already has a natural model to copy. In 2000, the type of vent he had predicted — alkaline and not too hot — was discovered[6]. The Lost City hydrothermal field lies in the Atlantic Ocean, 15 kilometres from the mid-ocean ridge. The vents pump out water that has been drawn down into cracks in the ocean floor and heated to about 200 °C. As the fluids return to the cold ocean, calcium carbonate precipitates out of the water, building 60-metre towers like gargantuan chemical gardens. And last year, oceanographers reported finding abiotically produced organic compounds, including methane, in the water flowing from the vent[7].

Russell thinks that his reactor might produce amino acids and peptides, but first he wants to test whether sulphide minerals standing in for the ocean crust will dissolve in the alkaline hydrothermal solution. That would be the initial

step toward the formation of the iron-sulphide chimneys that he believes provided a home for life's first metabolizing system.

It's a high-risk strategy, says Hazen, because the early Earth would have had many sources of organic molecules, and many places where they could have become concentrated. He likens Russell's approach of testing just one detailed possibility to starting a game of twenty questions by asking "is it Winston Churchill?" rather than "is it a man?". It is glorious if you're right, but it doesn't narrow things down much if you're wrong. "Mike Russell has a hunch, and there's nothing wrong with that," he says. "Maybe it'll be like winning the lotto and his hunch will be right, or maybe it wasn't Winston Churchill after all, and that leaves the other 6 billion people on Earth to go through."

A competing possibility is that one of the other four carbon-fixing pathways is a better candidate for the primordial biochemical reaction. Metabolism in acetogens and methanogens uses specialized enzymes, comments Eric Smith, a theoretical physicist and origin-of-life researcher at the Santa Fe Institute in New Mexico, and it is much more sensitive than other biological carbon-fixation pathways to changes in the isotope of carbon used. "That says that you're using a very carefully refined enzymatic reaction to do something difficult," he says. Instead, he and his colleagues believe that the initial biochemical pathway was the reductive citric-acid cycle. This pathway reverses the normal respiratory cycle seen in every oxygen-using cell, and some microbes still use it to fix carbon. It builds acetate, which has two carbons, up into citrate, which has six, and then breaks the citrate into acetate and oxaloacetate. This cycling can provide positive feedback that functions like reproduction, drawing more and more carbon into itself — an advantage that the one-way Wood–Ljungdahl pathway lacks.

Indoor hot springs

Smith thinks that the metabolism-first viewpoint is making headway, especially with the focus on hydrothermal vents. "The disagreements are really small compared with the basic orientation that we have in common," he says. He also thinks that the next move needs to be experimental; his colleagues are working to reproduce the reactions in the citric-acid cycle in vent-like laboratory conditions. Once someone gets their preferred pathway to work in the lab, "everyone can quiet down and say 'this is something we can agree on.'"

T. DAGAN

William Martin thinks that origin-of life research is 'unfalsifiable conjecture'.

Others, however, think that the whole metabolism-first theory is misconceived. The idea has two great flaws, says Steven Benner of the Foundation for Applied Molecular Evolution in Gainesville, Florida. In any system of organic reactions, some will make products other than those that might lead to life. "Organic chemistry has an intrinsic propensity to make tar," he says. "That tends to divert molecules out of any cycle." And any such set of reactions is unlikely to evolve greater complexity in a Darwinian fashion; instead it will just dissipate energy. What's more, he argues that acetyl CoA would not survive long at the temperature and alkalinity of Russell and Martin's favoured vent.

Benner aligns himself with the other main school of thought, which says that life started with a gene-like molecule that could catalyse its own replication. He has done many studies on how RNA could have first been made, although he also acknowledges the formidable difficulty of creating an RNA molecule large enough to behave like both a gene and an enzyme in abiotic conditions. The RNA camp did get a boost last week with a study suggesting it would be chemically easier to produce this molecule than was previously believed[8].

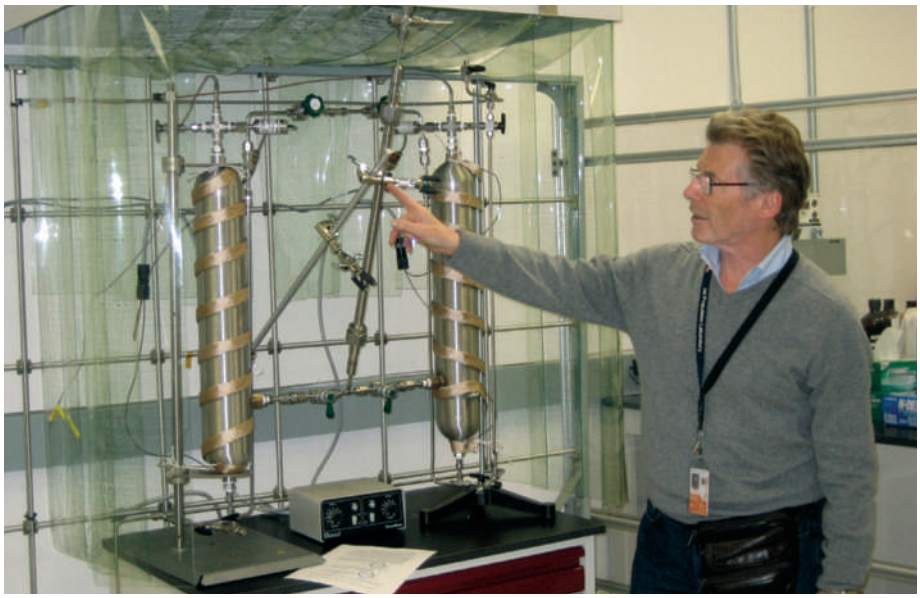
In the end, there is no agreement on how to solve the problem of life's origin. Martin thinks that research on the topic is "unfalsifiable conjecture" — the best we can hope for is a convincing story. "Even if you were to make a reactor in the laboratory, and put hydrogen and carbon dioxide and nitrogen in one end, and out pops something like *Escherichia coli* at the other end, you still couldn't prove that we and our ancestors arose that way. You'd just have a narrative that made it more plausible."

Russell thinks that if his reactor produces just about anything from tar to *E. coli* it will have been worthwhile — he quotes Thomas Edison's remark that he did not build 1,000 failed prototype light bulbs; rather he discovered 1,000 ways that the light bulb wouldn't work (see page 312 for one that did). Similarly, Russell hopes to help move the field forwards by sorting out what was possible and what was improbable around those warm vents, some 4 billion years ago. That's the most he can do, he says. "It's just a step at a time." ■

John Whitfield is a freelance science writer based in London and author of *In the Beat of a Heart: Life, Energy, and the Unity of Nature*.

1. Larter, R. C. L., Boyce, A. J. & Russell, M. J. *Mineralium Deposita* **16,** 309–318 (1981).
2. Fallick, A. E., Ilich, M. & Russell, M. J. *Econ. Geol.* **86,** 847–861 (1991).
3. Russell, M. J., Hall, A. J. & Turner, D. *Terra Nova* **1,** 238–241 (1989).
4. Martin, W. & Müller, M. *Nature* **392,** 37–41 (1998).
5. Martin, W. & Russell, M. J. *Phil. Trans. R. Soc. B.* **362,** 1887–1925 (2007).
6. Kelley, D. S. *et al. Nature* **412,** 145–149 (2001).
7. Proskurowski, G. *et al. Science* **319,** 604–607 (2008).
8. Powner, M. W., Gerland, B. & Sutherland, J. D. *Nature* **459,** 239–242 (2009).

NASA



Mike Russell is attempting to recreate the origins of life in his lab.

See News Feature, page 312.

CORRESPONDENCE

Calls to counter science scepticism are irrelevant in India

SIR — In the wake of Harry Collins's Essay 'We cannot live by scepticism alone' (*Nature* **458,** 30–31; 2009) and the Correspondence it stimulated from several Western scholars (*Nature* **458,** 702–703; 2009), I think it is also important to consider an Eastern viewpoint. Because of the notably different social set-up in India, Collins's call for studies to counter scepticism about science is irrelevant in this part of the world.

In an Indian context, my sense is that, in the significant segment of society for which such issues matter, science is neither the ultimate form of knowledge nor a victim of scepticism. Here, religion is the way of life. Even for many scientists and scholars of other disciplines, traditional religious values and philosophy are the unshakeable pillars in every domain of their lives, including science. Religion is a guard against the fear of future unknowns.

My observations as a research scientist of more than 30 years' standing suggest that most scientists in India conspicuously evoke the mysterious powers of gods and goddesses to help them achieve success in professional matters, such as publishing papers or gaining recognition. This is probably because factors outside their control come into play: religious endeavours offer comfort as well as being seen as a prerequisite for success.

In general, Indian society is not sceptical of science either — the common belief is that the boons of science outweigh any ill effects. After all, it has solved some of the toughest problems of humankind and has ushered in the era of technology-driven economies. It also addresses our curiosity and infuses a rational way of thinking into our societies. Acknowledging that uncertainty is an innate component of science should raise the standards and accuracy of scientific investigation, rather than increasing scepticism.

Lalit M. Kukreja Homi Bhabha National Institute and Laser Materials Processing Division, Raja Ramanna Centre for Advanced Technology, PO CAT, Indore 452 013, India
e-mail: kukreja@rrcat.gov.in

Protecting the environment can boost the economy

SIR — In attempting to sustain natural ecosystems, we should not assume that imposing a price on goods and services that adversely affect the environment will also have a negative effect on the economy. Placing a value on ecosystem services certainly changes the relative cost of various actions, but approaches being developed in other areas indicate that not all costs must necessarily rise.

Take the case of carbon emissions. Revenues from the sale of emissions permits to power-generating companies can be returned to the economy through funding of research into clean-energy technologies, say, or by returning money to the consumer — for example, under a 'cap-and-dividend' system that pays dividends to all taxpayers (whose environment is being damaged). This would alleviate the regressive nature of increased energy costs being passed on to consumers. For those purchasing 'green' power (wind, solar and hydro) generated with minimal carbon emissions, net costs would decrease.

Likewise, vehicle purchases could be governed by a 'feebate' system: those producing above-average emissions would cost more, the extra 'fee' being used to provide rebates to buyers of less-polluting vehicles.

There is no reason why ecosystem services could not be priced using comparable systems. For example, the fees paid for new development projects might be based on their environmental impact relative to some average, making developments cheaper if they can preserve valuable ecosystems or sequester carbon, and charging more to those that do not. Although goods and services having large adverse effects on the environment would increase in cost, those with minimal adverse effects would become relatively, or even absolutely, less expensive. And that, of course, is the whole point.

Drew Shindell NASA Goddard Institute for Space Studies, 2880 Broadway, New York 10025, USA
e-mail: drew.t.shindell@nasa.gov

Time for China to restore its natural wetlands

SIR — Your News story 'Putting China's wetlands on the map' (*Nature* **458,** 134; 2009) points out that almost 30% of China's natural wetlands vanished between 1990 and 2000. It is time for the country to restore these natural wetlands and, in view of their ecological importance, to construct some artificial wetlands to supplement them.

Wetlands include tidal marshes, mangroves, swamps and flood plains. They contribute the largest sector of total terrestrial ecosystem services — for example, flood mitigation, water-quality improvement, habitat biodiversity and landscape aesthetics. Their alarming rate of disappearance in China can be blamed on conversion to farmland and on pollution from point and non-point sources.

China is putting a massive stimulus package in place to boost the economy, covering the infrastructure of transportation, medical care, education and industrial upgrading. Water regulation is also a vital investment target, and will help to meet the increasing demand for water and create employment. The merits of water-related programmes in Western countries, such as best-management practices and low-impact development, and of small-scale rainfall-collection systems could be useful models for China to study.

Hydroelectric power and the south-to-north water-diversion project will help to alleviate drought in the future, as will China's investment in natural and constructed wetlands, and in additional water resources such as farmland irrigation and drainage.

Xubin Pan Department of Environmental Engineering, Texas A&M University Kingsville, Kingsville, Texas 78363, USA
Bin Wang State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
e-mail: wangbin07@gmail.com

Erasmus Darwin saw sexual selection before his grandson

SIR — I must leap to the defence of Charles Darwin's grandfather, Erasmus. It is not the case that, as David J. Hosken says in his Correspondence, "no one else envisaged anything like sexual selection" before Charles Darwin (*Nature* **458,** 831; 2009). His grandfather had done so more than 50 years earlier.

In his book *Zoonomia* (Johnson, 1794), Erasmus writes: "the three great objects of desire, which have changed the forms of many animals by their exertions to gratify them, are those of lust, hunger, and security". Lust, he goes on, leads to sexual selection: male birds fight so that "the strongest and most active animal should propagate the species, which should thence become improved".

Erasmus is often lost behind the glare of his stellar grandson, but he should not be forgotten. In many ways, as readers of *Zoonomia* and his great poem 'Temple of Nature' will appreciate, he played John the Baptist to Charles's Jesus.

C. U. M. Smith Vision Sciences, Aston University, Birmingham B4 7ET, UK
e-mail: c.u.m.smith@aston.ac.uk

COMMENTARY

Pandemics: good hygiene is not enough

The US government is doing well to communicate uncertainty over swine flu. It must also help the public to visualize what a bad pandemic might be like, says **Peter M. Sandman**.

By the time you read this, the outbreak of H1N1 'swine flu' may no longer seem to be a worldwide threat and the disease may have receded from the headlines. As the initial fuss dies down, public-health experts will remain on high alert, but the media and public will move on to something else, muttering about fear-mongering.

And whatever the situation is like now, it won't be the end of the story. A mutated virus (more virulent or transmissible or resistant to drugs) could appear a few months later.

As a risk-communication professional, I have been watching the US government walk a tightrope between over-reassurance and over-alarm about a swine-flu outbreak that could easily turn out to be devastating, relatively mild or anywhere in between. The United States hasn't issued false reassurances that they will keep the pandemic from 'our' shores — a temptation to which dozens of governments have succumbed. Here I will show what else I think the country is doing right — and wrong.

The US Centers for Disease Control and Prevention (CDC) is doing a superb job of explaining the current situation and how uncertain it is. The reiteration of uncertainty and what that means — advice may change; local strategies may differ — has been unprecedentedly good.

The CDC's biggest failure is in not doing enough to help people visualize what a bad pandemic might be like so they can understand and start preparing for the worst.

For the ordinary citizen, the US government has so far recommended only hygiene. It has told people to stay at home if they are sick and to wash their hands. It hasn't told people to stock up on food, water, prescription medicines or other key supplies. Two years ago in response to 'bird flu' worries, Mike Leavitt, the then US secretary of health and human services (HHS), was criss-crossing the country with that advice (www.pandemicflu.gov). Today, CDC officials won't say whether it is still good advice. It is.

> "Why are officials so wary of describing the worst case vividly and urging people to prepare for that?"

Richard Besser, the acting director of the CDC, isn't understating the risk. He says he is "very concerned", but expresses his concern with a soothing bedside manner. He doesn't have that rumpled, exhausted emergency-manager look that the Nuclear Regulatory Commission's Harold Denton perfected in the 1979 Three Mile Island crisis. Denton left people feeling that the risk was serious and that they were in good hands. Besser says it is serious but leaves us feeling that he doesn't want us to worry much.

Still, I don't fault Besser for looking and sounding reassuring. Good crisis communication means saying alarming things in a calm tone, and he is doing exactly that.

The problem is that he isn't giving us anything to do except being hygienic. He keeps telling us, accurately, that the CDC is being aggressive in its response to the outbreak. But he is not asking the public to take further action. He needs to urge citizens, schools, hospitals and local governments to follow Leavitt's advice.

Instead, we have a surreal situation in which the federal government has released one-quarter of the Strategic National Stockpile of antiviral drugs, so there will be enough oseltamivir (Tamiflu) to deploy to millions of sick Americans. But it hasn't yet asked those Americans to stock up on tinned fruit and peanut butter.

We've been here before. In 2005, the pandemic influenza threat came from avian H5N1. The CDC and HHS were similarly convinced that the risk was serious, similarly committed to aggressive preparatory action — that's why we have that Strategic National Stockpile — and similarly disinclined to alarm the public. The feeling was that people had been alarmed enough by the 11 September 2001 terrorist attacks and the wars in Afghanistan and Iraq, and that the government had exhausted its quota of scary utterances. There is much the same feeling today about the economic meltdown.

I was in the minority then, as I am now, urging officials to involve the public in pandemic-preparedness efforts. In early 2005 my recommendations fell largely on deaf ears.

Hygiene is useful, but getting ready for a pandemic also requires stocking up on key supplies.

R. LARSEN/THE GRAND RAPIDS PRESS/AP

Don't panic!

That summer, President George W. Bush read about the 1918 pandemic in John Barry's *The Great Influenza*. Then Hurricane Katrina hit New Orleans. The two together convinced the White House that raising concerns about worst-case scenarios was more appropriate than confident over-optimism. Soon the CDC and HHS were sounding the alarm about a possible pandemic. They aroused some concern, but no panic; they inspired some individual and community-preparedness efforts. And then attention shifted elsewhere, until now.

Why are officials so wary of describing the worst case vividly and urging people to prepare for that possibility? There are two reasons — first, a fear of fear itself. Although crisis-management experts have known for decades that panic is rare (http://tinyurl.com/ogofyw), officials routinely expect the public to panic if told alarming things, and misdiagnose orderly efforts to prepare as panic.

This approach nearly always backfires. Officials terrified of creating panic make over-reassuring statements, suppress alarming information and belittle those who are frightened as 'irrational'. Frightened people are left alone with their fears, persuaded that their government has betrayed them. This increases public anxiety, which officials cannot channel into effective action because they have already delegitimized it. During the 2003 severe acute respiratory syndrome (SARS) outbreaks, for example, the Chinese government denied that Beijing was seeing SARS cases and SARS deaths. These false denials led to actual panic in Beijing.

Predicting deaths

To its credit, the CDC has not made over-reassuring statements, suppressed alarming information or belittled people's fears. For several days before the first US swine-flu death on 29 April, Besser predicted that there would be US deaths. That is excellent risk communication. He has not understated how bad things were or how bad things could get. His failure has been subtler than that: sending the message that the CDC will do whatever it takes to protect us, and that we need do little or nothing to protect ourselves. From the outset, CDC messaging has aimed to keep us calm.

The second reason for the wariness of officials is a fear of being seen to overreact. Critics are already accusing officials of over-warning the public. And if the virus recedes and a pandemic never materializes, these critics will consider themselves proved right — as if the fact that your house didn't burn down this year proved the foolishness of last year's decision to buy insurance against fire. The only consolation I can offer officials is that many more people have lost their jobs for failing to take a disaster seriously than for being excessively alarmist about a possible disaster that never happened.

The risk-communication solution to this quandary is to issue warnings that are both scary and tentative. Public-health officials need to use the same sound bite to say, "This could get very bad, and it is time to prepare in case it does", and "This could fizzle out, and we'll probably feel a bit foolish if it does".

It might help if officials had a better understanding of the relationship between taking precautions and fear. Leaving aside the practical benefits, there are two psychological impacts worth describing.

First, consider the people officials are most worried about — those who are excessively alarmed. Here is a secret of preparedness that is easy to forget: it is calming to prepare. Having things to do gives people a sense of control. It builds confidence, and it makes them more able to bear their fear.

Second, there are those who are not worried, or who have already 'switched off'. Each time officials repeat practical advice, more people take it. Some of them take it sceptically, but take it nonetheless. Whenever someone acts, the scepticism is reduced. So urging people to prepare can calm those whose concern is excessive and rouse those whose concern is insufficient. It also offers the practical benefits of putting key supplies to hand.

As Besser says, we are currently in a "pre-pandemic" phase. The World Health Organization raised the alert level up from phase 3 to 4 on 27 April; and ratcheted it up again to phase 5 on 29 April. Phase 6 is a full-blown pandemic.

In announcing phase 5, Margaret Chan, the WHO director-general, echoed the CDC advice. When asked what individuals could do to protect themselves and their families, she advised hygiene and social distancing; wash your hands, stay home when sick, less hugging in public. But the WHO's own guidance for phase 5 emphasizes that a pandemic is "imminent" and that the time to finalize preparations is short. That ought to mean more action than reducing hugging.

We may stay at phase 5 for weeks or months. Or we could progress to a full pandemic that is mild, not catastrophic, or the threat could recede. So the key issue is what to say to the public when a pandemic seems imminent, but no one knows how it will turn out.

Two years ago, my wife and colleague Jody Lanard and I tried to answer that question in an online article. To aid officials we delineated 25 specific messages (see 'Things to say when a pandemic seems imminent') and the risk-communication rationales behind them.

Fundamentally, officials need to ask themselves whether they see the public as potential victims to be protected and reassured, like young children, or as pandemic fighters — grown-ups — who can play an active part in the crisis that might be ahead. The difference in tone could save lives. ■

THINGS TO SAY WHEN A PANDEMIC SEEMS IMMINENT

- It looks as if a flu pandemic is starting.
- It is no longer about the birds.
- This is a new warning, more urgent than any warning so far.
- The experts still aren't sure.
- We don't know how bad it will be.
- Here's what we know so far about the severity issue.
- Society will survive, but the pandemic may be very bad.
- We might have a window of opportunity now to make some practical preparations. Make the most of it — even though the effort might be wasted.
- What matters most is how households, neighbourhoods, community groups and businesses prepare.
- Individual and community preparations will focus on three tasks — reducing each person's chance of becoming sick, helping households with basic survival needs and minimizing and coping with larger societal disruption.
- Social distancing to impede contagion will be important but unpleasant.
- School closings present a difficult social-distancing dilemma.
- Hand-washing is far from a panacea. But it is easy, it is under your control and it has no significant downside.
- Like washing your hands, wearing a face mask may help a bit. But doing this has more downsides.
- Getting ready for a pandemic is largely about preparing for possible shortages.
- It is probably too late to stockpile much now, but do what you can.
- Now is the time to think about how to care for loved ones at home.
- To get ourselves through the hard times that may be coming, we will need volunteers. How can you help?
- If the pandemic is severe, the hardest job won't be coping with the disease. It will be sustaining the flow of essential goods and services, and maintaining civil order.
- Here's what the government is doing ...
- Try not to switch off. Try not to overreact.
- Even though we hope riots, panics and other sorts of civil disorder will not be common, it is important to be on guard.
- We are going into this pandemic crisis determined to be candid. That means you need to expect bad news, confusing changes in policy, conflicting opinions and conflicting information.
- Listen to stories about what 1918 was like, and to guesses about what the coming pandemic may be like.
- This is how to get more information ...

Adapted from a 2007 article (http://tinyurl.com/r6g6ur) by Peter M. Sandman and Jody Lanard.

Peter M. Sandman is a risk-communication consultant, 59 Ridgeview Road, Princeton, New Jersey 08540-7601, USA.
e-mail: peter@psandman.com

See also Essay, page 324, and for ongoing coverage of the H1N1 outbreak: www.nature.com/swineflu.
A longer version of this article is available at http://tinyurl.com/prbwf2.

ESSAY

Pandemics: avoiding the mistakes of 1918

As bodies piled up, the United States' response to the 'Spanish flu' was to tell the public that there was no cause for alarm. The authority figures who glossed over the truth lost their credibility, says **John M. Barry**.

In the next influenza pandemic, be it now or in the future, be the virus mild or virulent, the single most important weapon against the disease will be a vaccine. The second most important will be communication. History has shown that to cut vaccine production time, minimize economic and social disruption, deliver health care and even food, governments need to communicate well — both between themselves and with the public.

The US response to the 1918 flu offers a case study of a communication strategy to avoid. The world response to the threat of an emerging flu in recent weeks shows that we have learned from the past. And there is much to learn.

The pandemic that began in January 1918 and ended in June 1920 killed an estimated 35 million–100 million people worldwide, or 1.9–5.5% of the entire population[1]. Although an estimated 2% of people died in Western countries, some large subgroups were affected disproportionately. The Metropolitan Life Insurance Company, based in New York, found that the disease killed 3.26% of its insured US industrial workers aged 25–45. Given that 25–40% of the population contracted the disease, case mortality would have been 8–13% in that population[2].

> "The communication strategy of either reassurance or silence had its effect. Its effect was terror."

The flu started slowly. In the United States, a small wave of the disease sputtered across the country in the spring of 1918, but went largely unnoticed except in military training camps. The effects were more noticeable in Europe, where many soldiers in the armies of the First World War fell ill. By the end of summer, a more lethal wave had surfaced in Switzerland. On 3 August, the US military received an intelligence report comparing the Swiss epidemic to the Black Death.

The US government used the same strategy for communicating about the disease that it had developed to disseminate war news. The essence of that strategy was described by its main architect, writer Arthur Bullard: "Truth and falsehood are arbitrary terms … There is nothing in experience to tell us one is always preferable to the other … The force of an idea lies in its inspirational value. It matters very little if is true or false." Fellow adviser Walter Lippman, another architect of this strategy, sent President Woodrow Wilson a memo saying that most citizens were "mentally children" and advising that "self-determination" had to be subordinated to "order" and "prosperity". In 1917, the day after receiving Lippman's memo, Wilson issued an executive order to control all government communication strategy during the war that was premised on keeping up morale.

As a result, when the full-blown and lethal pandemic wave arrived in the United States in September 1918, Wilson never made a single statement about it, and lesser public figures provided only reassurance. US surgeon general Rupert Blue declared: "There is no cause for alarm if proper precautions are observed." Local health officials echoed this message. Chicago's director of public health, for instance, decided not to "interfere with the morale of the community", explaining: "It is our job to keep people from fear. Worry kills more than the disease."

That last phrase became a mantra repeated in hundreds of newspapers. Paid advertising carried a comparable message: every day, press advertisements for Vicks VapoRub appeared with the line: "Simply the Old-Fashioned grip masquerading under A New Name."

Yet it was not ordinary influenza by another name. The disease was unusual enough to be misdiagnosed initially as cholera, typhoid and dengue. Some people died within 24 hours of the first symptom. The most horrific feature was bleeding, not just from the nose and mouth but also from the ears and eyes.

Nonetheless, the government and newspapers continued to reassure. Although physicians fully understood the explosive nature of the pandemic, they routinely misled people, covered up the truth and lied. In Philadelphia, for example, public-health director Wilmer Krusen promised — before a single civilian had died — to "confine this disease to its present limits. In this we are sure to be successful." As the death toll grew, he repeatedly reassured the public that "the disease has about reached its crest. The situation is well in hand." When the number of daily deaths broke 200, he again promised: "The peak of the epidemic has been reached." When 300 died in a day, he said: "These deaths mark the high-water mark." Ultimately, daily deaths reached 759. The press never questioned him.

Meanwhile, the bodies piled up. In many cities, they lay uncollected in homes for days. In some places, including Philadelphia, they were buried in mass graves dug by steam shovels.

Same old fever?

Unfortunately, Philadelphia's communication strategy was the rule, not the exception. Local officials and newspapers across the country were either deceptive or said nothing. Many papers did not print lists of the dead. Even as 8,000 soldiers were hospitalized in Camp Pike, Arkansas, over four days, the *Arkansas Gazette* in Little Rock, just a few miles away, maintained, "Spanish influenza is plain la grippe — same old fever and chills." This communication strategy of either reassurance or silence had its effect. Its effect was terror.

Lies and silence cost authority figures credibility and trust. With no public official to believe in, people believed rumours and their most horrific imaginings. A man living in Washington described the result: "People were afraid to kiss one another, people were afraid to eat with one another … It destroyed those contacts and destroyed the intimacy that existed amongst people … there was an aura of a constant fear that you lived through from getting up in the morning to going to bed at night."

Under that pressure, society first drifted, then threatened to fall apart. The health-care system, already drained of physicians and nurses by the military, collapsed first. Elsewhere, fear, not illness, kept people at home. Absenteeism reached extraordinary levels. Shipyard workers were told that their duties were as important as a soldier's; they were paid only if they worked; and, unlike elsewhere, physicians were available to them on site. Yet absentee rates in the shipyards — one of the few industries for which there are good data — still ranged from 45% to 58% (ref. 3). Absenteeism crippled the railroad system, which transported nearly all freight, bringing it to the point of collapse. It shut down telephone exchanges, closing off communication, and further isolating and alienating people. Grocers refused to open. Coal sellers closed. In cities and rural communities, the Red Cross reported that people "were starving to death not for lack of food but because the well were too panic stricken to bring food to the sick".

Victor Vaughan, a sober, serious scientist,

SAN FRANCISCO CHRONICLE, TUESDAY, OCTOBER 22, 1918

WEAR A MASK
and Save Your Life!

The Emergency That Now Confronts Our City
Is Beyond the Facilities of the Health Department

The RED CROSS

has come to the assistance of the Board of Health. Doctors and nurses can not be obtained to take care of the afflicted. You must wear a mask, not only to protect yourself but your children and your neighbor from influenza, pneumonia and death.

A GAUZE MASK IS 99% PROOF AGAINST INFLUENZA

Doctors wear them. Those who do not wear them get sick. The man or woman or child who will not wear a mask now is a dangerous slacker.

WEAR MASKS

GOING TO WORK
AT WORK
GOING HOME
AT HOME

Bold declarations like this one from San Francisco were rare in 1918.

for years dean of the medical school at the University of Michigan in Ann Arbor, worried that if this trend accelerated "for a few more weeks … civilization could easily disappear from the face of the Earth".

Better communication led to better results. In San Francisco, for example, despite a slow reaction to the initial onslaught of flu, in October 1918 the mayor, health officials and business and union leaders all signed a full-page newspaper advert in huge type reading: "Wear A Mask and Save Your Life!" It was a rare, bold statement. In this city, society, although reeling, functioned. Food was delivered, and the sick were cared for. Where people had accurate information and knew what they faced, they often performed heroically. Red Cross professionals, physicians and nurses routinely risked their lives. When Philadelphia's city police — who knew the facts even if the papers weren't printing them — were asked to supply four volunteers to "remove bodies from beds … and load them in vehicles", 118 officers responded[4].

Truth telling

Of course, the world is different today to how it was in 1918. But communication remains paramount.

Until we develop a vaccine that is effective against all influenza viruses and available globally, the world will be vulnerable to influenza pandemics. H1N1 is the most immediate danger, including the possibility that a more deadly wave will strike later this year, but H5N1 and other viruses remain pandemic threats.

First and foremost, authorities must tell each other the truth. This provides crucial lead time for vaccine production. Models even suggest that in a few circumstances, surveillance and transparency may allow a new virus to be contained and extirpated.

The world has performed well in the past few weeks in this regard, but this is a lesson that has not been entirely taken on board. In 2003, China initially covered up SARS, putting the world at risk and contributing to near-panic in Beijing, where people felt they could trust nothing coming from the government. In 2004, Thailand and Indonesia withheld information during the first outbreaks of H5N1 bird flu. There continue to be both political and bureaucratic problems in ensuring that H5N1 isolates are shared — especially by Indonesia — thereby increasing the risk to the world.

BETTMANN/CORBIS, *THE SAN FRANCISCO CHRONICLE*

Telling the public the truth is also paramount. Before a vaccine becomes available during a severe pandemic, at some point the government will ask citizens to adhere to a series of public-health guidelines for non-pharmaceutical interventions, such as staying at home if they become ill. Large-scale, sustained compliance will be essential if those measures are to succeed. Compliance requires trust, and that depends on truth-telling.

In the United States, former health and human services secretaries Tommy Thompson and Michael Leavitt deserve credit for institutionalizing real transparency in the current US pandemic plan. And the administration of President Barack Obama has performed admirably so far. Obama himself has addressed the issue several times, making perhaps only one mistake, when he said the threat was "cause for concern, but not alarm". Had things deteriorated quickly, he ran the risk of suddenly having to reverse his position.

In Mexico, the problem was not reticence but candour, releasing inaccurate information that overstated the problem. Mexico should be congratulated for this, not condemned.

Although a false alarm can be damaging, it is not nearly as damaging as silence — the type of silence that makes people believe the truth is being withheld. That is how trust disintegrates and how rumours — passed in the streets in 1918, today passed over Internet blogs — take hold and grow.

I don't much care for the term 'risk communication'. It implies that the truth is being managed. The truth should not be managed, it should be told. Only by knowing the truth can imaginary horrors be transformed into concrete realities. And only then can people start to deal with those realities, and do so without panic. ■

John M. Barry is a Distinguished Scholar at the Center for Bioenvironmental Research of Tulane and Xavier Universities, New Orleans, Louisiana 70112, USA. He is author of *The Great Influenza: The Epic Story of the Deadliest Plague In History*.
e-mail: jbarry@tulane.edu

1. Johnson, N. P. A. S. & Mueller, J. *Bull. Hist. Med.* **76,** 105–115 (2002).
2. Jordan, E. O. *Epidemic Influenza: A Survey* (American Medical Association, 1927).
3. Turner, C. E. *Report Upon Preventive Measure Adopted in New England Shipyards of the Emergency Fleet Corporation* Record Group 90, entry 10, file 1622, US National Archives.
4. Mayor's Annual Report for 1918, City of Philadelphia, 40.

See Commentary, page 322, and for coverage of the H1N1 outbreak: www.nature.com/swineflu. See http://tinyurl.com/fluessay for further reading.

BOOKS & ARTS

Paul Dirac: a physicist of few words

A detailed biography argues that the Nobel prizewinner's notorious reticence delayed experimentalists from discovering the antimatter that would confirm his elegant theory, explains **Frank Close**.

The Strangest Man: The Hidden Life of Paul Dirac
by Graham Farmelo
Faber & Faber: 2009. 560 pp. £22.50

BETTMANN/CORBIS

Paul Dirac predicted the anti-electron's existence, but did little to encourage others to hunt for it.

Among scientists, Paul Dirac is widely regarded as being in the same league as Albert Einstein. In London's Westminster Abbey, Dirac's eponymous equation describing the quantum behaviour of electrons is set in stone. But in his home town of Bristol, UK, his reputation is overshadowed by that of his fellow student at the Bishop Road School, Archie Leach — better known as the film star Cary Grant. On asking the Bristol Record Office for material about Dirac for his new book, author Graham Farmelo received the response: "Who?"

Danish physicist Niels Bohr described Dirac as "the strangest man". His extreme reticence, monosyllabic responses and repetitious statements are legendary. Six years elapsed before even close colleagues learned any of P. A. M. Dirac's forenames. When he came up with the equations of quantum mechanics, his weekly postcard home merely said, "Not much to report here." After solving a decades-old problem by creating Fermi–Dirac statistics later that year, the news once more was: "Not much to report." When he arrived at the relativistic quantum equation that describes the electron, he didn't even send a postcard. Even his colleagues were unaware of it.

In this elegantly written biography, Farmelo's meticulous research sheds considerable light on Dirac's personality and the circumstances behind it. Several members of Dirac's extended family developed acute depression, six committing suicide within a century, including his brother. His father was cold and authoritarian, his mother overweening — the description of her excruciating behaviour at the Nobel prize ceremony, haranguing journalists and officials on behalf of her idolized son, is pure entertainment. Dirac immersed himself in mathematics.

The received wisdom is that in producing his equation for the electron, Dirac 'discovered' the concept of antimatter in 1928, and four years later, Carl Anderson's discovery of the positron in cosmic rays validated Dirac's idea. But in Farmelo's account the reality is rather different.

Dirac's electron equation — declared "achingly beautiful" by physicist Frank Wilczek — described the spin of the electron, and caused a sensation once people began to understand its unusual structure. However, it contained puzzling solutions in which the electron had negative energy. Dirac proposed that a vacuum is filled with a sea of negative-energy electrons. Any hole in this vacuum would appear as a positively charged, positive-energy particle. At first, he thought that this particle was the proton, until J. Robert Oppenheimer pointed out that if this were so, the electron and proton could destroy each other and matter would be unstable. Wolfgang Pauli was equally sceptical, remarking that anyone making a theory of matter should first apply it to the atoms of their own body. Pauli went on to prove that the positive particle must have the same mass as an electron, which was worrying because experimenters had not found any such particle. With the debate unresolved, many began to wonder if Dirac's equation might be wrong.

In 1931, Dirac referred for the first time to the 'anti-electron', remarking that it could not occur in nature owing to its immediate destruction by ubiquitous electrons. Although he commented that it could be made transiently in experiments, he was surprisingly circumspect, more concerned with the difficulties of detection than the inevitability of its existence. He made no suggestion as to how experimentalists might make it, or recognize it. He was away in the United States later that year when Robert Millikan gave a talk at the University of Cambridge, UK, showing Anderson's images of particle tracks from cosmic rays — including some that looked like those of electrons but which curved the wrong way in a magnetic field. No one associated these tracks with Dirac's holes.

By 1932, the holes had become a joke. At a meeting in Copenhagen, when Bohr lost his patience and confronted Dirac with: "Do you believe all that stuff?", he simply replied, "I don't think anyone has put a conclusive argument against it." Dirac no longer seemed to be strongly committed to the anti-electron; the absence of the particle was, Farmelo says, "sapping his morale". He even told Werner Heisenberg that he had ceased to believe in it.

On 2 August 1932, Anderson found his first clear single particle trail, now hailed in textbooks as the 'discovery of the positron'. This realization was far from clear cut, however. In a series of missed opportunities, no one seemed able to put two and two together to link Dirac's holes and Anderson's 'positron'.

Anderson published his positron paper in September 1932 in the journal *Science*. But remarkably, no one in Cambridge seemed to have read it. By that autumn, British physicist Patrick Blackett had his own images of positrons, and had even shown them in a talk with Dirac and Soviet physicist Peter Kapitsa in the audience. Dirac stayed silent. Kapitsa exclaimed

"Now, Dirac, put that into your theory! Positive electrons, eh!" Farmelo comments that Kapitsa "had spent hours talking with Dirac but had evidently not even heard of the anti-electron" and that Dirac simply replied "Positive electrons have been in the theory for a very long time". Yet there is no sense that Dirac was claiming anything, apparently convinced that the positive trails in the pictures were "a mirage". Farmelo sees Dirac as exhibiting "reticence taken to the point of perversity". His colleagues so mistrusted his abstract theory that they could not accept that it predicted new particles.

The first link between hole theory and the positron came from Blackett, who showed sensational images of electron–positron pair creation at a meeting at the Royal Society in London, saying that they "fit extraordinarily well with Dirac's hole theory". Immediately afterwards, journalists rushed to interview him. Meanwhile Dirac, who was lecturing in another room in the same building, was "unavailable for comment".

According to Farmelo, Dirac later realized that he held responsibility for not having advocated that experimentalists should hunt for positrons, nor advising on how to detect them. Had he done so, the positron could have been discovered "in a single afternoon", as Anderson put it. When asked later why he did not speak out and predict the positron, Dirac said, "pure cowardice".

Nonetheless, Dirac on other occasions believed that he had predicted it, although not everyone agreed. Blackett said: "Dirac nearly but not quite predicted the positron." So much for history; today, Dirac's role in foreseeing the positron, and the mirror world of antimatter, was, as Farmelo describes it, "one of the greatest achievements in science".

Farmelo concludes *The Strangest Man* by analysing Dirac's singular character and genius. He makes a sound case that Dirac was autistic, and argues that his behavioural traits were crucial to his success as a theoretical physicist. Cambridge in the 1920s was the ideal environment for him: tolerant of eccentricity; college life providing for his every need; the rules of dining at High Table enabling a rigidly predictable form of social contact. These unusual circumstances enabled Dirac's special genius to flower. As to autism, this is thought to be caused by disrupted brain development, which can show up as irregularities in brain tissue. These can be visualized using positron emission tomography scans — the medical application of Dirac's antimatter. Irony indeed. ■

Frank Close is professor of physics and a Fellow at Exeter College, University of Oxford, Oxford, UK. He is the author of *Antimatter*.
e-mail: f.close1@physics.ox.ac.uk

The end of the invasion?

Invasion Biology
by Mark A. Davis
Oxford University Press: 2009. 288 pp. $55

Ascension Island in the South Atlantic Ocean is a good example of the changes that invasive species can wreak. Its volcanic mountain tops once hosted a monotonous carpet of ferns. But in 1843, botanist Joseph Hooker recommended that the bleak island be wooded by importing many new plants — what modern ecologists would see as a massive, human-mediated biological invasion. Surprisingly, this resulted not in ecological meltdown, but in the creation of a lush cloud forest. The forest traps mists, cycles nutrients and survives, generation after generation, without its species having evolved together. A study of this anomalous system is cited in Mark Davis's new book *Invasion Biology*. Why? Maybe because it is not so anomalous.

Invasion Biology starts out as a graduate-level text on how organisms brought far from their homes by humans can flourish, often at the expense of native species in the places they 'invade'. But on turning the pages, the book reveals itself to be an iconoclastic argument that much of the field's conventional wisdom is wrong, that biologists are more swayed by their emotions about invasive species than they care to admit, and that invasion biology as a field should be disbanded. Davis writes, "This may be the first time that an author has concluded a book, the title of which is the same as the discipline being reviewed, by recommending that participants consider abolishing their discipline."

Davis is not on the fringe. His arguments crystallize a rumbling of dissent recently heard among those who study invasive species. As he puts it, "There is little about biological invasions that make them so unique that a specialized subdiscipline need be sustained to study them."

Invasion biology began in earnest in 1958 when ecologist Charles Elton published his pioneering book, *The Ecology of Invasions by Animals and Plants* (see *Nature* **452,** 34; 2008). Elton saw species 'invasions' in the context of niches. In an intact, co-evolved ecosystem, every species will have a slightly different role, or niche, and often every niche will be filled. For example, predators eat herbivores; herbivores eat plants; some plants grow on wet soil and some grow on dry. When new species are introduced, the theory goes, they can get a foothold only by finding a vacant niche or by throwing out another species.

Niche theory gives rise to the diversity-invasibility hypothesis, which posits that the more species there are in an ecosystem, the more niches will be filled and the harder it will be for a new species to become established.

But the evidence does not bear this out. Many studies have failed to find any strong relationship between how diverse a place is and how easy it is to invade. Davis concludes that, despite its appeal and its "implicit affirmation of the value of diversity", the hypothesis is not true. In fact, the opposite may hold. In any ecosystem, each individual plant or animal has to get a foothold, irrespective of its origin. A seed does not care whether it is exotic or native when it lands on the ground, and neither do the surrounding species. The key insight is that there is nothing fundamentally different about exotics other than where they came from.

K. SCHAFER/ALAMY

Ascension Island: not all imported species are destructive.

Davis challenges other received wisdom, such as the idea that newcomers are more likely to compete with or predate on natives than help them flourish, and that introduced populations are unlikely to be genetically diverse. He refuses to exaggerate the differences between natives and exotics, or to see exotics as the enemy.

Elton's 1958 book was an expansion of a series of radio broadcasts aimed at the public. Davis speculates that this audience was the reason behind Elton's colourful, militaristic comparisons of "ecological explosions" with bombs. This may have sown the seeds of the current 'good-versus-evil' rhetoric of species invasion, with its talk of biological pollution, killer weeds and battling garlic mustard.

Davis is not a fan of such heated rhetoric. He feels that the dichotomous approach is not ecologically enlightening. Life is much messier, more dynamic and more complex, he says. He stuffs the book with examples of exotic species

that play nicely with their new neighbours. For every pest, there are many more unobtrusive immigrants that live quietly in their new haunts, even helping the growth and development of native species. This does not mean that invading species are never a problem, but Davis argues that they are not always troublesome.

Davis writes well, and clearly. But his big contribution is to the sceptical re-examination of the field as a whole. This book will not kill it off. But if, over time, invasion biology were to become absorbed into broader ecological fields that focus on the movement of species, future historians of science might see *Invasion Biology* as the beginning of the end. ■

Emma Marris writes for *Nature* from Missouri.

Paper ambassadors of science

Philip Parker of Britain's Royal Mail celebrates special stamps and his new set for the 250th anniversary of the Royal Botanic Gardens at Kew.

After Queen Elizabeth II, the most featured individual on British postage stamps is a scientist. Charles Darwin has appeared on four stamp issues, in 16 different stamp designs, in the past 30 years. The last set, issued on 12 February this year to commemorate the bicentenary of his birth, used a jigsaw design to illustrate the interconnectedness of the varied disciplines — zoology, botany, geology — that Darwin synthesized into his theory of evolution. A separate sheet of four stamps makes up the hydrographic map of the Galapagos Islands that resulted from the voyage of HMS *Beagle*.

Darwin's popularity as a subject for stamps is appropriate because he was a prodigious letter writer. From its introduction in 1840, the Penny Post was the Internet of its day, facilitating peer review among scientists. With the service came the postage stamp, which is arguably the most widespread and visible platform for public art.

For 50 years, alongside the everyday stamps showing the Queen, Royal Mail has been issuing pictorial stamps to mark aspects of British heritage and contemporary life. They are produced in hundreds of millions, and competition for topics is fierce. Every year Royal Mail receives around 2,000 requests for subjects. These are filtered using certain criteria — anniversaries are covered in 50-year multiples, and themes must be of national importance or celebrate the national character. Intensive desk research and public consultation funnels these down to a continuing programme of around 13 or 14 stamp issues per year.

Subjects are chosen for a range of audiences, from postal historians to the average letter sender. Themes include both light and shade; for example, an impressive set on the grandest cathedrals can be followed by stamps celebrating the fiftieth birthday of the *Carry On* comedy films — and both can be equally successful.

Postal offices worldwide issue pictorial stamps, and science is frequently celebrated. Scientific concepts are often difficult to illustrate concisely, so scientists are more often depicted. The handsome 2008 stamps from the United States feature portraits of chemist Linus Pauling and astronomer Edwin Hubble, among others.

Indeed, one of the most popular stamp-collecting themes globally is astronomy. But flora and fauna are consistently attractive to the public, and science and engineering topics generally do well. Royal Mail's earliest special stamps highlighted the opening of the Forth Road Bridge in 1964, and Jodrell Bank's radio telescope in 1966. However, success is all in the detail and in the translation of the subject on to the tiny canvas of a stamp.

Stamps are arguably the most visible form of public art.

ROYAL MAIL GROUP

Published on 19 May, the Royal Mail's latest stamp issue marks the 250th anniversary of the founding of the Royal Botanic Gardens at Kew, near London. A set of four stamps feature images of key landmarks at both Kew and Wakehurst Place in West Sussex, where the Millennium Seed Bank aims to conserve 10% of the world's seed-bearing flora by 2010. Alongside are ten stamps of UK endangered plants, many of which Kew is actively conserving. Delicate botanical art is used to portray these species, six examples of which are drawn from Kew's extensive art collection.

Royal Mail works closely with partner organizations, such as Kew, and other experts in the field being portrayed, to cross-check every fact and ensure the content is accurate. In-house specialists commission and manage the work of external designers and illustrators, who may work on the same subject but to different briefs. One specialist might consider photography, another might create new illustrations, and a third could explore existing botanical art. The preferred approach is picked after consultation and discussion with the independent Stamp Advisory Committee. For the ten plant stamps, the style of botanical art was found to give the clearest, most accurate and most engaging depictions in such a small space. Once the final designs are proofed to satisfaction, they are submitted for approval by the Queen before printing.

The Plants sequence is the latest in the Action for Species series. Every year, ten stamps are issued depicting threatened UK species, for which there are conservation plans in place. The series began with Birds in 2007, continued with Insects in 2008, and a set on mammals is being prepared for next year. The series has been devised as a countdown to the International Year of Biodiversity in 2010. Another major set of forthcoming stamps will mark the Royal Society's 350th anniversary in 2010.

Stamps are 'paper ambassadors'. Affixed to letters and parcels, they can end up in any corner of the world, where the receiver will form an opinion of the sender, of the country of origin, that country's sense of self and its global contribution. This is why Royal Mail pays so much attention to detail on its stamps — they illustrate, in the best sense, the best of British. ■

Philip Parker is head of stamp policy at Royal Mail, 35 Rathbone Place, London W1T 1HQ, UK.
e-mail: philip.parker@royalmail.com

For details of Royal Mail's Plants stamps, see online at www.royalmail.com/plants.

Q&A: The exhibition designer

A pioneer of interactive museum installations, **Edwin Schlossberg** lets young visitors experience science first hand, from launching a space shuttle to seeing the world through an animal's eyes. As his neuroscience-inspired paintings are shown this month in New York City, he explains how he applies cognitive science to harness children's curiosity.

How do you engage children with science?
If you put a bucket of water in front of a child — 2 years old, 5 years old, even 8 years old — they will play with it forever. They learn a lot because they can craft a range of experiences as they integrate their sensory and physical worlds. I try to design like that. Most science museums try to train future scientists or to say "Isn't science cool?" To me, neither of those attitudes is appropriate. I like to make experiences that allow you to see something differently, in a way that encourages you to have a conversation with other people in the room. It's more about provoking questions than giving answers.

S. WILKES

What challenges do science museums face?
Today's parents are afraid of their kids growing smarter than them. When the theme park Sesame Place opened in Dallas, Texas, in the early 1980s, a survey found that the vast majority of parents would not come because they were worried their children would ask questions they couldn't answer. They were afraid of their kids' curiosity. We decided to print tens of thousands of comic books that answered all the questions kids might ask. And we got an audience.

How do you draw on cognitive science?
I want to make exhibits that engage all the senses, so I look to people who have the best understanding of neurophysiology and learning. When I designed for the Brooklyn Children's Museum in New York, I talked with child psychologist Jean Piaget. Then I read sociologist Erving Goffman's *The Presentation of Self in Everyday Life*, which argues that we often behave like actors, taking on roles that influence how we respond to our surroundings and each other. I've consulted educational computer scientist Seymour Papert and artificial-intelligence pioneer Marvin Minsky at the Massachusetts Institute of Technology. Most of these scientists think that learning by doing is better than just looking or hearing.

How does the public see neuroscience?
Amazing discoveries are happening in labs all over the world, but they are not visible to the public. I'm not aware of any current major exhibit on neuroscience in the United States or abroad. For the US pavilion at the 2005 World's Fair in Japan, I proposed an exhibit with my company ESI Design on how people are imaging neurons. We thought it would be important, beautiful and interesting. But President George W. Bush didn't like it.

You designed a museum for NASA at the Stennis Space Center in Mississippi?
Yes, it's right next to the highway — the first science museum in the country that is also a rest stop. It is now scheduled to open next year because the site was completely destroyed by Hurricane Katrina and construction was delayed. It's a space museum, but the main focus will be on meteorology. You'll walk into a big spherical theatre, put on 3D glasses and feel like you're at the centre of a hurricane. Then the sides of the theatre will roll up to reveal labs where you can explore the tools that allow us to make the weather visible. You'll be able to turn on sensors to monitor the wind speed and water temperature at buoys in the Gulf of Mexico, and compare that with what you see outside the building.

What other exhibits are you working on?
Did you ever see a Tamagotchi, the digital toy that would 'die' if you didn't pay attention to it? I found that idea creepy but brilliant. For the exhibit at the Children's Museum of Los Angeles in California [which was due to open next year, but the funding for which is now uncertain], we tried to make it feel as if it was an ecosystem the children had to take care of with their own hands. They walk into this fantastical place with a giant tree and animals called Dogbear and Puppycub that seem to be sleeping. If the kids start to blow air and shine lights, the creatures wake up and they can pet and feed them.

Is there anything you've always wanted to build?
An oversize scale model of the human body as a giant pinball game. It would be the size of an American-football field. A hundred people playing together would make all the systems work so the body wakes up. It might help us to think of ourselves not just as individuals, but as a gigantic community of cells.

You're also a painter. How does your art relate to the brain?
My new set of paintings shows what I imagine your neuron patterning would be if you were thinking of a phrase — such as "You being focused", "You considering stillness", "You absolutely certain". It thrills me that scientists are able to see neurons. My art is what we might see if we could witness the process of thinking itself. ■

Interview by **Jascha Hoffman**, a writer based in New York.
e-mail: jascha@jaschahoffman.com

Edwin Schlossberg: At the Moment
Ronald Feldman Fine Arts, New York City
Until 30 May 2009.

NEWS & VIEWS

NANOTECHNOLOGY

Another dimension for DNA art

Thomas H. LaBean

Many of nature's intricate nanostructures self-assemble from subunits. Efforts to mimic these assembly processes enter a new phase with a method to design and build three-dimensional DNA nanostructures.

Through the ages, some of the most iconic and lasting artefacts of human ingenuity have been sculptures and carvings, created from a wide variety of materials. But until now, a general-purpose material from which nanometre-scale, three-dimensional shapes could be made has been lacking. On page 414 of this issue, Douglas *et al.*[1] introduce a clever method for fabricating nanometre-scale objects from DNA, and report the construction of several such objects. The authors describe their method as "analogous to sculpture from a porous crystalline block", except that the structure of their block consists of tubes — DNA double helices — arranged in a regular honeycomb lattice. The desired shapes are not literally carved into the starting material, but instead form from DNA that has been designed to self-assemble into a supramolecular complex.

The use of DNA as a construction material for making nanometre-scale objects began more than 25 years ago[2], and has since developed into the field of structural DNA nanotechnology[3,4]. The field relies on the fact that molecular recognition and assembly of DNA can be programmed so that it forms designed nanostructures. Such programming is enabled by our understanding of Watson–Crick base pairing: for any DNA base sequence, we can immediately determine the complementary sequence, and know that the two molecules will find and bind to one another in water under appropriate conditions. Well-developed synthesis techniques allow DNA strands of any desired base sequence to be easily prepared.

In 1998, DNA nanotechnology was transformed by the introduction of the 'tile and lattice' strategy[5]. Tiles are nanometre-scale building blocks that fold independently, and typically contain domains of DNA double helices tethered by 'strand-exchange points'. These points model naturally occurring junctions that form in genetic-recombination complexes when DNA strands are traded between helices. Short, single-stranded DNA segments hang off the tiles at strategic locations. On cooling in solution, the single-stranded segments on different tiles bind to each other, so that the tiles assemble into larger, predominantly two-dimensional lattices.

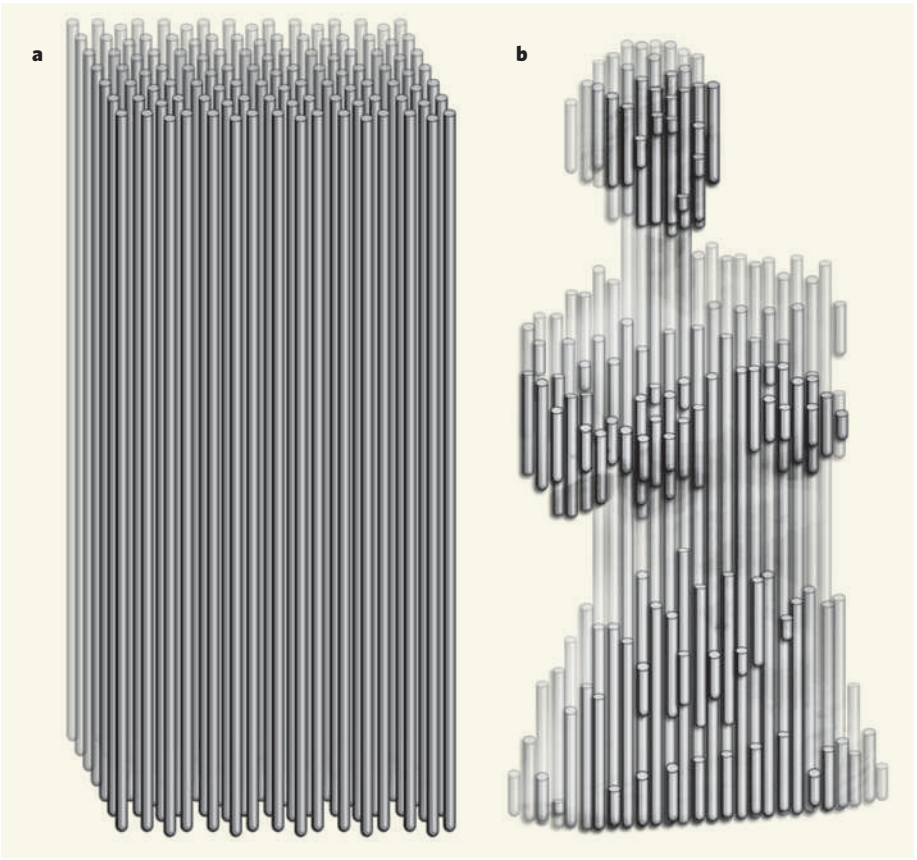


Figure 1 | DNA sculpture. Douglas *et al.*[1] report a method for designing and constructing three-dimensional nanostructures from DNA. **a**, The computer-aided design process begins with a block of tubes arranged in a honeycomb lattice. **b**, A template for the desired DNA structure is designed by removing sections of the tubes, just like carving a sculpture from a block. The remaining tubes will become DNA duplexes in the final object. The DNA structure is designed by routing a single-stranded scaffold DNA (a virus genome) through every section of the tube template. Hundreds of short strands of DNA are then designed to bind to the folded scaffold, cross-linking between different tubes and 'stapling' together the overall structure. When the staple molecules are synthesized and mixed with the scaffold DNA in solution under appropriate conditions, they direct the folding of the scaffold into the desired nanostructure. The structure shown here is more complex than those prepared by the authors (see Fig. 2 on page 416).

Another conceptual leap occurred in 2006, with the demonstration of DNA 'origami'[6]. This strategy uses a long, single-stranded DNA — for example, the genome of the M13 virus — as a scaffold molecule that is folded back and forth on itself to form a planar raft of double helices. The resulting structure is knitted together by a few hundred short, synthetic DNA molecules that act as staples, linking together the helices at appropriately spaced strand-exchange points. The raster-like routing of DNA scaffolds through origami structures provides a general system for making nanometre-scale, two-dimensional sheets of any shape, and with any desired surface pattern. But given the flat raft architecture,

it is not easily used to make intrinsically three-dimensional objects.

Douglas and colleagues' work[1] represents a third revolution in DNA nanotechnology. They have extended the DNA origami technique by showing how a DNA scaffold strand can form layers of helices arranged in a honeycomb lattice, thus providing a general-purpose, crystalline material from which three-dimensional objects can be constructed. In principle, any shape can be made from this DNA material, as long as it can be 'carved out' from a block of the honeycomb lattice.

To design their nanostructures, the authors devised a computer-aided process that begins with a template block composed of tubes (Fig. 1); each tube becomes a DNA duplex in the final structure. Once a target shape has been defined by removing sections of the block, a single-stranded scaffold DNA (the M13 virus genome, as in flat DNA origami) is routed through every part of the structure, and complementary 'staple' strands are designed to bind to the scaffold and thus create duplexes. Finally, strand-exchange points are defined between neighbouring double helices. Enough of these junctions must be used to stabilize the overall structure, while still maintaining enough flexibility in the system to allow the desired shape to assemble. Having drawn up plans for their target structures, Douglas *et al.*[1] heated, then very slowly cooled, a solution of the scaffold DNA and its hundreds of staples. Under these conditions, the staples directed the folding of the scaffolds into the desired shapes.

Douglas and colleagues' approach can be compared with a recently published procedure for three-dimensional DNA origami[7], in which a hollow box ($42 \times 36 \times 36$ nanometres) was assembled. Two-dimensional DNA origami was used to construct all six flat walls of the box on a single scaffold strand, and then inter-wall staple strands directed the assembly of the final three-dimensional form. The box design is highly innovative — it even includes a lid that can be opened and closed — but the box gains its three-dimensionality by orienting intrinsically two-dimensional subunits against one another in space. By contrast, the honeycomb lattice technique[1] is inherently three-dimensional from the start of the design process.

Of course, the primary goal of DNA nanotechnology is not to create aesthetically pleasing sculptures, but to make functional devices and materials. For practical applications, structures generated using Douglas and colleagues' method will probably need to be integrated with other nanomaterials that have electronic, photonic or catalytic properties superior to those of DNA. There are currently also other limitations to the technique. For example, the self-assembly process results in low product yield (providing only about 7–44% of the theoretical yield), proceeds very slowly (taking about a week), and generates products that have an unfavourably high charge density (because the charged DNA backbone is packed tightly in space). Furthermore, the upper limits on the total size of the products and the lower limits on their feature resolution have yet to be determined. The shapes that have been made so far are also somewhat blocky (see Fig. 2 on page 416); the sculpture depicted in Fig. 1b of this article would require either a larger scaffold strand than is currently used, or several such strands.

Nevertheless, the potential of Douglas and colleagues' technique is clear. Hierarchical structures, constructed from several repeating subunits, are a much-sought-after goal of nanotechnology, and the authors present three examples in their paper, including a stunning icosahedron assembled from three M13 genome scaffolds (see Fig. 4 on page 418). This successful move into three dimensions heralds a new era for the field of structural DNA nanotechnology. ■

Thomas H. LaBean is in the Departments of Computer Science, Chemistry and Biomedical Engineering, Duke University, Durham, North Carolina 27708, USA.
e-mail: thomas.labean@duke.edu

1. Douglas, S. M. *et al. Nature* **459,** 414–418 (2009).
2. Seeman, N. C. *J. Theor. Biol.* **99,** 237–247 (1982).
3. Seeman, N. C. *Nature* **421,** 427–431 (2003).
4. LaBean, T. H. & Li, H. *Nano Today* **2,** 26–35 (2007).
5. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. *Nature* **394,** 539–544 (1998).
6. Rothemund, P. W. K. *Nature* **440,** 297–302 (2006).
7. Andersen, E. S. *et al. Nature* **459,** 73–76 (2009).

COMPUTATION

The edge of reductionism

P.-M. Binder

Research at the frontier between computer science and physics illustrates the shortcomings of the reductionist approach to science, which explains macroscopic behaviour using microscopic principles.

In his 1972 paper "More is different", Philip Anderson[1] claimed that multi-component physical systems can exhibit macroscopic behaviour that cannot be understood from the laws that govern their microscopic parts — a feature known as emergent or complex behaviour. Anderson's position is at odds with that of Stephen Hawking, who once suggested[2] that, as soon as all fundamental laws of the Universe are understood, we will in principle be able to explain all macroscopic phenomena. Writing in *Physica D*, Gu and colleagues[3] provide a beautiful illustration of a physical system that cannot be easily 'reduced', and of the developing symbiosis between theoretical physics and computer science[4].

To address 'the understandable', Stephen Wolfram[5] examined the relation between computation and the unfolding of the physical world. He defined as reducible those systems for which there is a computational shortcut that allows their behaviour to be efficiently predicted rather than reproduced step by step. For example, the motion of a simple pendulum is described by a cosine function that can be computed using a rapidly converging mathematical series, rather than simulating each and every pendulum oscillation. Such shortcuts do not usually exist for chaotic systems, for example.

Wolfram made an additional, important point. Many systems are irreducible, but among them only a few are undecidable: they have properties that cannot be formally calculated, as stated in Kurt Gödel's and Alan Turing's theorems[6]. Undecidability is a property of universal computers or Turing machines. Macs, PCs and DNA computers[7] with unlimited memory would qualify as such machines. And this is where the notion of 'different' (or complex) systems can be made more precise — those with undecidable global properties despite having well-understood local (microscopic) governing laws.

As a first example of undecidability, consider a cellular automaton (CA) — a lattice of cells, each of which can take on a finite number of values (states) and evolves over time according to the configuration of a set of neighbouring cells. This is the microscopic transition rule. For the one-dimensional CA known as 'elementary rule 110', two states are allowed ('0' or '1'), and any cell will evolve to 0 if either its state and that of its right-neighbour cell are 0, or if its state and those of both its immediate neighbours are 1 — otherwise it will evolve to 1. Thus, the local governing law is fully understood.

But the global dynamics of a CA is a different matter, as can be seen in Figure 1. Each row displays the lattice at a different time step, thus providing a full spatiotemporal record of the dynamics of the system. Cells far apart act in concert to sustain 'particles'[8]: structures that move and interact, and in doing so, compute. The result is an intricate and undecidable global dynamics.

It is not easy to demonstrate that rule 110 can simulate a universal computer[9]. Such proofs often involve the construction of a few logic gates and information channels that allow universal computation to be implemented, and could well be argued to be reductionist. But once these elements have been constructed, the step that shows that a system has undecidable

DAVID J. ECK

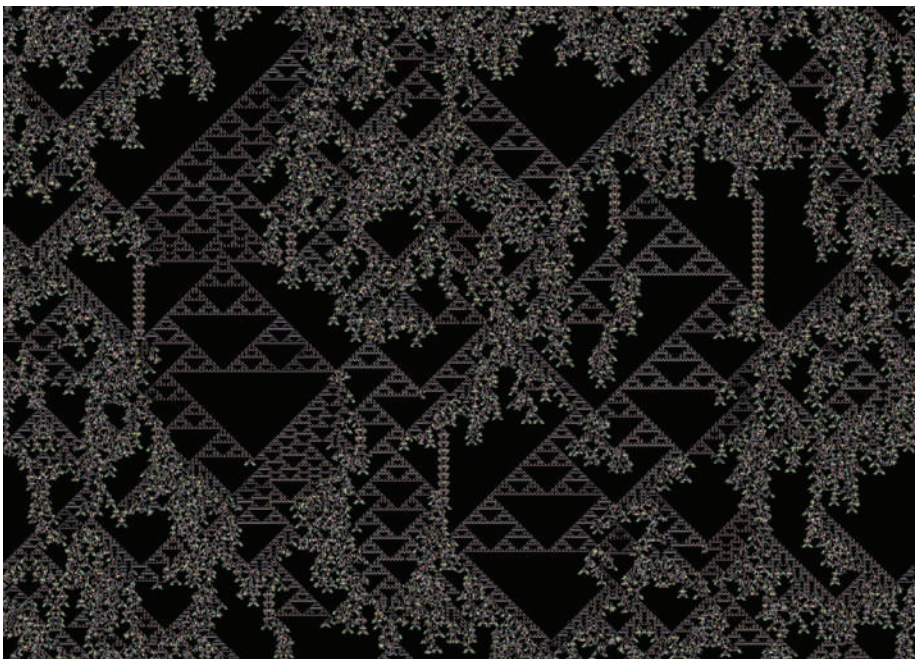


Figure 1 | Evolution of a cellular automaton. A cellular automaton is a lattice of cells that evolves through a number of time steps according to a predefined rule, which stipulates the next state of a cell on the basis of its current state and that of its immediate neighbours. In this example, each row represents the same lattice, 1024 pixels across, at a different time step, for a total of 768 steps. The colour of each pixel denotes one of the eight possible states of the cell. The initial condition (top row) was randomly chosen. As time proceeds, patterns ('particles') can be identified that move to the left or right and interact with each other, leading to complex and formally undecidable global dynamics. Gu and colleagues' findings[3] are based on the interpretation of this image as a two-dimensional spatial array of spins at a fixed time.

properties involves proof by contradiction rather than constructive proof: a higher level of abstraction. When Gu *et al.*[3] write "the understanding of macroscopic order is likely to require additional insights", they may have in mind procedures such as proofs by contradiction that transcend mere reductionism.

The authors focused on the Ising model: a lattice of spins that interact with one other and with an external magnetic field. The individual spin states can be 0 or 1 (corresponding to 'up' or 'down' magnetization), just like those of elementary CA. The main difference is in the dynamical rule: spins tend to align with their neighbours (and with the external field, if one is applied to the system), whereas thermal fluctuations counteract and randomize their state. Therefore, the microscopic transition rules are probabilistic. Ising models in more than one dimension exhibit phase transitions: at sufficiently low temperatures, the tendency of spins to align overcomes thermal jiggling, and the system becomes and remains ordered. Perhaps not surprisingly, the physics and mathematics immediately around the disorder-to-order phase transition are rich, and have been well studied.

In their study, Gu *et al.*[3] mapped the dynamics[10] of a certain CA into the lowest-energy (ground) states of Ising models. In this framework, Figure 1 can now be interpreted as a snapshot of a two-dimensional spatial lattice of spins. They grouped spins into blocks that encode the logic operations needed to produce universal computation in the corresponding CA. They then defined the 'prosperity', p, of two-state systems as "the probability that a randomly chosen cell at a random time step is live" (live meaning state 1).

Using the computational properties of the CA, Gu and colleagues were able to show that p is undecidable for infinite, periodic Ising systems. They argued that, as a consequence, many macroscopic properties of an Ising system, including the system's magnetization and degeneracy (number of independent configurations) at zero temperature, depend on p and hence are also undecidable. Because Ising models have been used to describe not only magnetic materials but also neural activity, protein folding and bird flocking, the consequences of Gu and colleagues' results transcend both computer science and physics.

Alas, their results apply only to infinite lattices, and hence seem of limited use. The finite Turing systems one would encounter in real life are decidable. But there are hints that finite objects may, after all, have undecidable properties. One hint comes from certain mappings of a solid square onto itself, which have been shown to be undecidable[11,12]. These procedures slice and rearrange parts of the square in a way that allows computer operations such as shifts to be implemented, and they take advantage of real numbers (which require an infinite number of digits) to pack an infinite computer into a finite region. A second hint comes from a new level of computation[13] that is

50 YEARS AGO

The Neutrino. By Prof. James S. Allen — This small book gives an excellent description of experimental work on the neutrino, which only in the past few years has been shown to have any direct physical property apart from balancing energy, momentum and spin in β-decays. It is a tribute to modern methods of experiment that neutrino-capture in hydrogen has been demonstrated by Reines and Cowan in spite of the fact that a neutrino beam could traverse a thickness of solid material measured in *light years* without appreciable loss ... It has not happened often that so tricky a field of physics has been cleared up so quickly, as a result of brilliant theoretical work, which suggested many key experiments.
From *Nature* 23 May 1959.

100 YEARS AGO

The Problem of Age, Growth, and Death: a Study of Cytomorphosis. By Prof. Charles S. Minot — From the time of Cicero, perhaps before, the problems of longevity and of the cause of old age have again and again been subjects of speculation. Not long ago, Metchnikoff, in his optimistic work, "The Nature of Man," ascribed old age to poisoning by bacterial poisons developed as a result of fermentations occurring in the large intestine ... Prof. Minot develops another conception of the nature of "growing old." Although in old age a condition of atrophy is frequent, and various degenerations of cells and tissue are usually present, in particular of the arterial system, so that it has been said "a man is only as old as his arteries," Prof. Minot combats the view that old age is a kind of disease, and regards it as a necessary consequence of the changes in the cells of the body, which are inevitably progressive from birth to death; this succession of cellular changes is termed "cytomorphosis."
From *Nature* 20 May 1909

more powerful than a Turing machine, and has been proposed as just the right one to simulate natural physical phenomena. One hopes that the work of Gu *et al.*[3], along with these two ideas, will lead to a better understanding of the 'computer' in which we live. ■

P.-M. Binder is in the Department of Physics and Astronomy, University of Hawaii, Hilo, Hawaii 96720, USA, and the Kavli Institute for Theoretical Physics, University of California, Santa Barbara.
e-mail: pbinder@hawaii.edu

1. Anderson, P. W. *Science* **177,** 393–396 (1972).
2. Hawking, S. W. *Is the End in Sight for Theoretical Physics?* (Cambridge Univ. Press, 1980).
3. Gu, M., Weedbrook, C., Perales, A. & Nielsen, M. A. *Physica D* **238,** 835–839 (2009).
4. Percus, A., Istrate, G. & Moore, C. (eds) *Computational Complexity and Statistical Physics* (Oxford Univ. Press, 2006).
5. Wolfram, S. *Phys. Rev. Lett.* **54,** 735–738 (1985).
6. Binder, P.-M. *Nature* **455,** 884–885 (2008).
7. Benenson, Y., Gil, B., Ben-Dor, U., Adar, R. & Shapiro, E. *Nature* **429,** 423–429 (2004).
8. Toffoli, T. & Margolus, N. *Cellular Automata Machines* (MIT Press, 1987).
9. Cook, M. *Complex Systems* **15,** 1–40 (2004).
10. Domany, E. & Kinzel, W. *Phys. Rev. Lett.* **53,** 311–314 (1984).
11. Moore, C. *Phys. Rev. Lett.* **64,** 2354–2357 (1990).
12. Bennett, C. H. *Nature* **346,** 606–607 (1990).
13. Siegelmann, H. T. *Science* **268,** 545–548 (1995).

SYSTEMS BIOLOGY

When it is time to die

Philippe Bastiaens

Why do cells of the same population respond differently to external death-inducing stimuli? Individuality seems to originate from non-genetic differences in the levels and activation states of proteins.

Any cell biologist can tell you that individual cells from a clonal cell population respond differently to the same stimulus, some not responding at all. In such cases the percentage of responders is seen as a measure of the experimenter's control over parameters that affect the stimulus, such as uniformity of the cellular environment. Variability in cell response can have grave implications. For instance, some tumour cells refuse to die in response to drugs that trigger programmed cell death (apoptosis), affecting the efficacy of chemotherapy. In this issue, Spencer *et al.*[1] (page 428) show that the non-uniform response of a human cell population to the apoptotic stimulus TRAIL can be ascribed to an intrinsic random factor: the naturally occurring differences in protein-expression levels.

To induce apoptosis, TRAIL binds to the cell-surface receptors DR4 and DR5, triggering specific intracellular signalling pathways. These receptors are therefore attractive targets for the development of anticancer drugs, and several compounds that can activate them have been tested in preclinical and in phase I clinical trials, with some promising results[2]. DR4 and DR5 are expressed in normal as well as cancerous tissues, although there is some indication that tumour cells might have higher levels of these receptors[2]; at least, the compounds tested in the trials selectively induce apoptosis in tumour cells. Nonetheless, significant problems remain, including resistance and differences in sensitivity to TRAIL, and fractional killing — situations in which successive cycles of chemotherapy kill only some of the tumour cells[3].

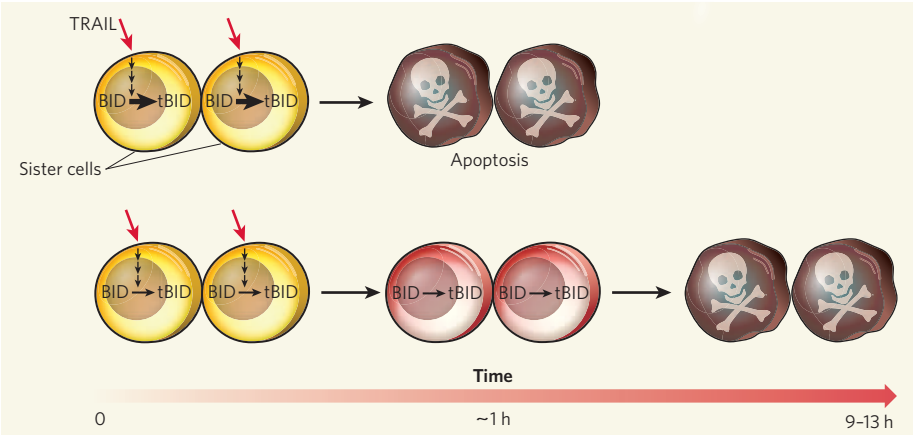


Figure 1 | Non-genetic factors contribute to the rate of apoptosis. Spencer *et al.*[1] find that different sets of sister cells respond differently in the time they take to die after exposure to the apoptotic stimulus TRAIL. The rate of response to TRAIL depends on the rate of proteolytic conversion of the pro-apoptotic protein BID to its active form tBID, which itself is affected by variance in the expression levels of several proteins in the early apoptotic machinery (vertical black arrows). The protein levels are inherited by daughter cells (which become neighbouring sister cells) causing them to behave similarly in response to TRAIL.

Spencer *et al.*[1] show that, in a cancer cell line, TRAIL induces a non-uniform response: some cells die within 45 minutes, some 8–12 hours later, and yet others do not die at all. Intriguingly, following exposure to TRAIL, recently born sister cells die after a similar period of time, suggesting that variability in the population arises from inherited cell differences before treatment with TRAIL (Fig. 1). The authors also find that, on inhibition of protein synthesis, the 'sisterhood' memory persists for longer, an observation that relates a non-genetic factor — protein expression — to variability in cell responses. Finally, they use computer simulation of a biochemical reaction model for apoptosis[4]. The stimulation used as input differences in the levels of proteins mediating apoptosis and the range in 'death times' the authors detect using this method match those they observed experimentally[1].

To investigate the molecular basis of variable cellular responses to TRAIL, Spencer and colleagues grouped the apoptotic protein machinery into three tiers: those occurring before, during and after the process of mitochondrial outer-membrane permeabilization (MOMP), which is crucial for apoptosis. In the first reaction tier, TRAIL binds to its receptors and leads to their association; the death-inducing signalling complexes (DISCs) assemble; and the proteolytic initiator-caspase enzymes become active to trigger MOMP. In the second reaction tier, during MOMP, mitochondrial proteins such as cytochrome *c* and SMAC are released into the cytoplasm. There, they activate effector caspases in the post-MOMP third reaction tier, causing cell death. The authors could microscopically image the activity of these mediators from each of the three tiers in single cells with genetically encoded fluorescence indicators[5] for both caspases and MOMP.

They find that variability in the time to death was almost exclusively determined by differences in the reaction rate of the initiator caspases, which convert a pro-apoptotic protein called BID into its truncated active form (tBID) in the first reaction tier (Fig. 1). tBID then induces the self-assembly of two pore-forming proteins, BAX and BAK — an activity that is normally prevented by its interaction with the anti-apoptotic proteins of the BCL2 family — into mitochondrial pores, thereby initiating MOMP. The authors therefore conclude that time to death is set by the rate of approach to a threshold in the levels of activated tBID at which mitochondrial pores form.

Spencer *et al.* argued that the levels of proteins functioning in the first reaction tier (DR4, DR5, DISC components, the initiator caspases 8 and 10, and BID) should determine the reaction rate for BID activation. But the authors' computer-simulation data show that the level of any one protein in the first tier does not determine time to death. Only on increased expression of one of these components did the

computer simulations detect a clear relationship between protein level and time to death — a finding that the authors verified experimentally by overexpressing a fluorescent BID protein. They therefore conclude that, under normal conditions, variations in the expression of several proteins together control the rate of approach to MOMP and so cell death.

Does this mean that several different states of the protein network in the first tier give rise to the same rate of approach to MOMP? To relate network states to time to death, the correlation between multiple protein levels and time to death must be investigated, rather than correlations involving single proteins as studied by the authors (using simulations).

Spencer and colleagues' work for the first time relates variability in molecular processes determined by protein-expression levels to variability in phenotypic response in human cells. However, it is not clear how the short-term memory ($t_{1/2}$ = 1.5 hours) for the rate of approach to MOMP can account for fractional killing in a cell population that has a typical doubling time of 20 hours. If protein levels fluctuate during this short time window, then successive TRAIL administrations should eventually kill all cells because of the high probability that the cells will experience the protein-network state favouring rapid apoptosis before they divide. Thus, optimized TRAIL-administration schemes could be devised to maximize the killing of tumour cells. Nevertheless, the ultimate cell-fate decision, to die or not to die, might eventually lie in the expression level of anti-apoptotic proteins, such as those of the BCL2 family, that set the threshold in the system. ■

Philippe Bastiaens is at the Max Planck Institute of Molecular Physiology, 44227 Dortmund, Germany.
e-mail: philippe.bastiaens@mpi-dortmund.mpg.de

1. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. *Nature* **459,** 428–432 (2009).
2. Ashkenazi, A. *Nature Rev. Drug Discov.* **7,** 1001–1012 (2008).
3. Newsom-Davis, T., Prieske, S. & Walczak, H. *Apoptosis* **14,** 607–623 (2009).
4. Albeck, J. G., Burke, J. M., Spencer, S. L., Lauffenburger, D. A. & Sorger, P. K. *PLoS Biol.* **6,** e299 (2008).
5. Albeck, J. G. *et al. Mol. Cell* **30,** 11–25 (2008).

EARTH SCIENCE

Life battered but unbowed

Lynn J. Rothschild

Early in its history, Earth experienced a pounding from extraterrestrial impacts. But instead of sterilizing the planet, it allowed microbial life to persist, according to numerical models of Earth's crust.

The great baseball player Leroy 'Satchel' Paige, who pitched professionally until the alleged age of 60, asked "How old would you be if you didn't know how old you were?". It was a reasonable question, given that his year of birth was unknown. Much the same applies to the age of life on Earth. If you are comfortable thinking that we Earthlings must be younger than 3,900 million years, the paper by Abramov and Mojzsis[1] on page 419 of this issue will come as a shock. We may be several hundred million years older than that.

The current narrative of Earth's history begins about 4,600 million years (4.6 Gyr) ago when the planet formed. Within 100 million years, the young Earth suffered its worst day ever with the impact of another — rogue — planet, the debris from which ultimately formed the Moon. Even if Earth had cooled substantially before the lunar-forming impact, the heat of the impact would have melted Earth's surface[2]. For another 600 million years, especially during the late heavy bombardment (LHB) — roughly 4.1–3.9 Gyr ago — Earth was pummelled by impactors (Fig. 1), mainly asteroids, that heated up both land and oceans. Few, if any, rocks remain from before the end of the LHB, with the oldest currently known being from 4.03 Gyr ago[3], and with the earliest, albeit controversial, isotopic fossil evidence for life at 3.8 Gyr ago[4].

Geological models[5] have suggested that Earth could not have been continuously habitable during this period, because there was insufficient time for the origin of life between impacts. Furthermore, estimates[6] of the last impact that sterilized the oceans range from 3.8 to 4.44 Gyr ago. This left biologists with the tantalizing but impenetrable possibility that life originated more than once, and that the present lineage of all life — the prokaryotic archaeans and bacteria, and eukaryotes such as ourselves — stems from a later incarnation. This view implies that life arose quickly, because there is isotopic evidence for its origin within 100 million years of the end of the LHB. Alternatively, life on Earth originated elsewhere, Mars for example.

NASA/JPL/USGS

Figure 1 | Impact evidence on the pock-marked face of the Moon. The late heavy bombardment is thought to have affected all bodies in the inner Solar System. On Earth, traces of the impacts have been erased by surface geological processes. But on the Moon many craters date to that time, and are still visible.

This story began to change with the discovery of zircons that pre-date the LHB. Zircons are telltale mineral inclusions in ancient rocks, and the evidence from them supported the idea of a Hadean era (4.38–3.85 Gyr ago) that was far from hellish, with liquid water, crustal recycling, a granitoid crust and low-temperature processes occurring at the boundaries of tectonic plates[7,8].

What if the period since the lunar-forming impact, including the LHB, was never severe enough to sterilize Earth? This is precisely what Abramov and Mojzsis[1] suggest. On the basis of numerical models of impact-generated heat in Earth's crust, they propose that, at worst, only 37% of Earth's surface was sterilized, and less than 10% experienced temperatures above 500 °C. They conclude that, even if all of the LHB impacts occurred simultaneously, Earth would not have been sterilized. The likelihood of finding life on planets with an equally strong impactor record, such as Mars, or maybe even extrasolar planets, has just improved.

The study[1] is based on several assumptions. The first is that life was distributed over the surface and subsurface of Earth by events during the LHB. Of course, there is no evidence for this, but today organisms up to 1–2 millimetres in size are globally dispersed by wind, oceanic circulation and groundwater networks[9], processes that would also have occurred

during the LHB. Transport of small particulates was more likely at that time, because a moist surface is more likely to retain particles, and impacts during the LHB vaporized any water and made the atmosphere rich in steam. With fluid movement through pore spaces in rock, microbes could reach a depth of about 10 kilometres within a thousand years[10].

The second assumption is that Earth's habitable zone extends to 4 km below the surface. Present microbial biomes exist in basalt rocks at 2.8 km depth with no apparent reliance on photosynthetically produced materials[11], and certain bacteria have been obtained from a 3.5-km borehole, but are thought to live 5.3 km or more below the surface[12]. The third assumption is that life can flourish at 110 °C (which may be a conservative estimate given the report[13] of an archaean organism that thrives at 121 °C; and who knows about extinct life?). Finally, there is the assumption that sterilized areas will be recolonized, which is reasonable given the processes outlined above.

All that said, what are the implications of Abramov and Mojzsis' conclusion? The most obvious is that it allows for a much earlier origin of life, by hundreds of millions of years. Moreover, it opens the possibility that life arose on Earth only once, and that the planet has been continuously inhabited ever since.

Molecular data[14–16] imply that the last universal common ancestor, or LUCA, was a thermophile, an organism that thrives at high temperatures (although there is evidence[17] that thermophily arose independently in the Archaea and the Bacteria). If the first life form was a thermophile, Abramov and Mojzsis' model suggests that life could have arisen during the LHB, or before then while Earth was cooling after the lunar-forming impact. But the model is evidence against the possibility that the LHB was a bottleneck traversed only by thermophilic prokaryotes, because life disperses quickly. Although some habitable areas, particularly for mesophiles, would have been destroyed for periods of time, new hydrothermal habitable areas would have been created. But at no point would all habitats for mesophiles have disappeared. What seemed an unknowable prehistory is slowly taking shape.

Even with the new study, we are left with the possibility that the photic (sunlit) zone of Earth's surface could have been destroyed intermittently throughout the LHB, as posited by Sleep *et al.*[6]. If so, it is unclear how secure life would have been in the absence of all possibility of photosynthesis. It might not be too difficult for photosynthesis itself to evolve[18], however, and it could conceivably have originated multiple times between impacts that sterilized the photic zone.

So, maybe we are an elderly 4.2-plus Gyr. In the year of Darwin mania, take comfort in the words of the other bicentenarian, Abraham Lincoln: "In the end, it's not the years in your life that count. It's the life in your years." Earthlings, we are still going strong. ■

Lynn J. Rothschild is at the NASA Ames Research Center, Moffett Field, California 94035-1000, USA.
e-mail: lynn.j.rothschild@nasa.gov

1. Abramov, O. & Mojzsis, S. J. *Nature* **459,** 419–422 (2009).
2. Nisbet, E. G. & Sleep, N. H. *Nature* **409,** 1083–1091 (2001).
3. Bowring, S. A. & Williams, I. S. *Contrib. Mineral. Petrol.* **134,** 3–16 (1999).
4. Whitehouse, M. J., Myers, J. S. & Fedo, C. M. *J. Geol. Soc.* **166,** 335–348 (2009).
5. Maher, K. A. & Stevenson, D. J. *Nature* **331,** 612–614 (1988).
6. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. *Nature* **342,** 139–142 (1989).
7. Wilde, S. A., Valley, J. W., Peck, W. H. & Graham, C. M. *Nature* **409,** 175–178 (2001).
8. Mojzsis, S. J., Harrison, T. M. & Pidgeon, R. T. *Nature* **409,** 178–181 (2001).
9. Finlay, B. *Science* **296,** 1061–1063 (2002).
10. Gold, T. *Proc. Natl Acad. Sci. USA* **89,** 6045–6049 (1992).
11. Lin, L.-H. *et al. Science* **314,** 479–482 (2006).
12. Szewzyk, U., Szewzyk, R. & Stenström, T.-A. *Proc. Natl Acad. Sci. USA* **91,** 1810–1813 (1994).
13. Kashefi, K. & Lovley, D. R. *Science* **301,** 934 (2003).
14. Pace, N. R. *Science* **276,** 734–740 (1997).
15. Di Giulio, M. *J. Theor. Biol.* **221,** 425–436 (2003).
16. Schwartzman, D. W. & Lineweaver, C. H. *Biochem. Soc. Trans.* **32,** 168–171 (2004).
17. Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N. & Gouy, M. *Nature* **456,** 942–946 (2008).
18. Rothschild, L. J. *Phil. Trans. R. Soc. B* **363,** 2787–2801 (2008).

CANCER

Melanoma troops massed

Paul H. Huang and Richard Marais

In many cancers, regulation of specific signalling molecules goes awry, affecting a host of other proteins and cellular processes. Proteomics is a useful systemic approach for identifying such extensive effects.

In most Western countries, the incidence of melanoma — a form of skin cancer linked to exposure to sunlight — doubles roughly every decade. Early removal of melanoma lesions leads to a good prognosis, but the outlook for patients with advanced melanoma is often bleak owing to a lack of effective treatments. To develop better therapies, a deeper understanding of this cancer's biology is needed. To this end, a paper published by Old and colleagues[1] in *Molecular Cell*, which reports on a new use of a proteomics technique to identify molecules regulating melanoma-cell behaviour, is a welcome advance.

A central player in melanoma is the signalling protein BRAF, which through its protein-kinase activity phosphorylates other proteins, regulating their activity. In normal melanocytes (pigmented skin cells from which melanoma develops), the membrane-bound protein NRAS activates BRAF, which then activates a second protein kinase, MEK. In turn, MEK activates yet another protein kinase, ERK (Fig. 1). This signalling pathway controls diverse cellular behaviours including growth, survival, differentiation and senescence. In at least half of melanoma cases BRAF is mutated, activating this signalling pathway independently of NRAS and so driving tumour maintenance and progression[2] (Fig. 1).

A key question is how mutant BRAF selects growth and survival in melanoma over the other cellular responses stimulated by the normal version of this protein. To answer this, it is necessary to identify the proteins that BRAF regulates in melanoma, and this is exactly what Old *et al.*[1] set out to do using a mass-spectrometry-based technique called negative ion scanning. This technique allows accurate characterization of many phosphorylated proteins without the need for sample purification, thus overcoming a central problem inherent in 'phospho-proteome' studies.

In melanoma cells with a mutation in BRAF, the authors identify a total of 568 phosphorylation sites on various proteins. To pinpoint which phosphorylation events are regulated by mutant BRAF, they treat the cells with a MEK inhibitor, narrowing their list to 90 phosphorylated proteins. Reassuringly, several of these proteins are known ERK targets. Nonetheless, Old and colleagues' list contains new substrates, among them FAM129B — a previously uncharacterized ERK substrate. Using an *in vitro* model system that is thought to mimic the way that melanoma cells spread through patients' bodies, the authors demonstrate that BRAF can stimulate melanoma-cell invasion by inducing FAM129B phosphorylation by ERK. Previous work[3] has indicated that melanoma cells are highly invasive and can spread to distant sites, contributing to the aggressive nature of this disease.

As expected, many of the other targets Old *et al.* identify seem to be phosphorylated directly by ERK, by ERK-related protein kinases or by kinases that ERK regulates (Fig. 1). Intriguingly, unquenched BRAF signalling also suppresses phosphorylation of 30 substrates in melanoma cells[1], suggesting that these proteins are indirect targets of this pathway that are regulated by complex feedback loops or crosstalk between different pathways. Furthermore, the authors' screen identifies many proteins with acidic phosphorylation motifs — sequences that are considered excellent potential target sites for the protein kinase CK2 (ref. 4). CK2 is linked to cancer-cell growth[4], a finding that places it downstream of BRAF in melanoma cells. Some studies, however, have placed CK2 upstream of BRAF and the related protein kinase CRAF[5]. It is possible,

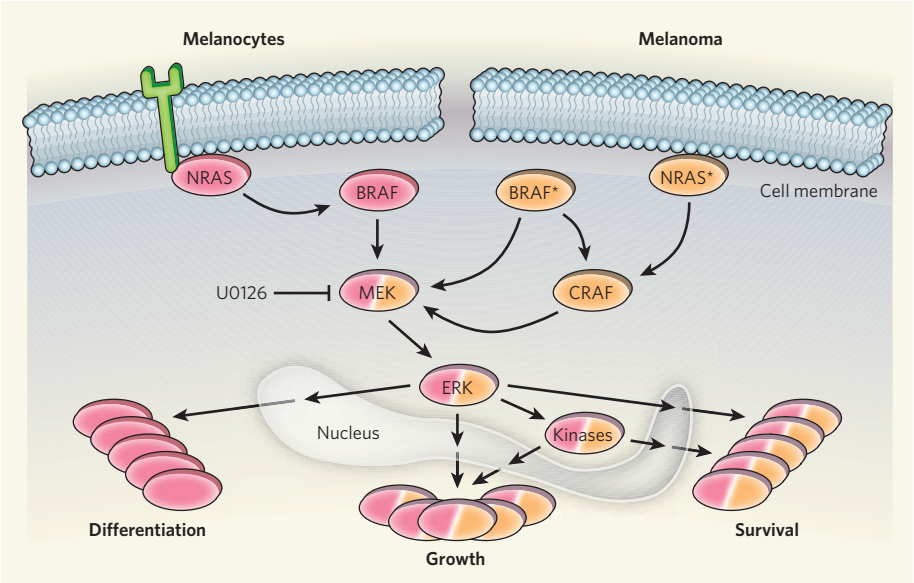


Figure 1 | The BRAF-mediated pathway in health and cancer. In healthy melanocytes, the NRAS–BRAF–MEK–ERK signalling cascade (pink) tightly regulates cellular functions such as differentiation, growth and survival. In melanoma (orange), BRAF mutations (BRAF*) bypass activation by NRAS, leading to cancer-associated signalling through the MEK–ERK pathway that favours growth and survival over differentiation. BRAF* can also activate this pathway through direct activation of CRAF. Mutant NRAS (NRAS*), however, activates MEK–ERK independently of BRAF, through CRAF. Old *et al.*[1] use the MEK inhibitor U0126 to identify the MEK signalling signature in melanoma using a mass-spectrometry-based technique.

therefore, that there is a positive-feedback loop between BRAF/MEK signalling and CK2 in melanoma cells.

It is intriguing that, whereas Old and colleagues' paper indicates that mutant BRAF regulates a complex network of substrates and pathways, 20 years of intense research[2,6] have suggested that the only direct substrates of BRAF are MEK and CRAF, and that BRAF signals through the MEK-mediated pathway alone. Unfortunately, because Old *et al.*[1] used a MEK inhibitor rather than a BRAF inhibitor, their powerful approach might have failed to identify direct BRAF substrates.

Furthermore, NRAS is notably mutated in about 20% of melanoma cases[2], and mutant NRAS activates MEK and ERK — albeit independently of BRAF — through CRAF[7] (Fig. 1). Again, because the 'BRAF signalling signature' Old *et al.* identify relies on MEK inhibition, it could be that some or perhaps all of the phosphorylation changes they detect will also occur as a result of NRAS mutations in melanoma cells — something the authors did not test. We therefore suggest that this study[1] reveals a signalling signature of MEK rather than of BRAF, and that some or perhaps all of the molecular targets identified will be common to all melanomas in which this pathway is abnormally activated.

Although many components are likely to be missing from this signature[1], which provides only a snapshot of the cancer networks regulated by MEK in BRAF-mutant cells, it is nevertheless tempting to use these data to gain insight into how BRAF rewires cell signalling. However, this possibility must be tested directly by comparing this signature with that characteristic of normal BRAF activity in melanocytes. It would also be appropriate to repeat this analysis using a second, and possibly more potent, MEK inhibitor that could reveal any off-target, drug-specific responses.

The power of Old and colleagues' approach is evident from their identification of both the involvement of a previously uncharacterized protein (FAM129B) in melanoma-cell invasion, and new and potentially essential interactions between BRAF and other signalling molecules such as CK2. To extend their data, other sophisticated methods must be used. For example, the chemical-genetics approaches pioneered by Shokat and colleagues[8] — whereby protein kinases are tailored to use modified ATP molecules as a source of a phosphate group — should make it possible to determine which kinases phosphorylate which distinct substrates.

Undoubtedly, by confirming the substrate–kinase connections that Old and colleagues' data suggest, future studies will enrich our understanding of BRAF-mediated signalling in melanoma. This work therefore forms an exciting framework to better understand melanoma-associated changes. Perhaps more importantly, it provides an approach for the identification of targets and network connections that will ultimately serve as a rich resource to design more effective therapies for melanoma. ■

Paul H. Huang and Richard Marais are at the Institute of Cancer Research, London SW3 6JB, UK.
e-mail: richard.marais@icr.ac.uk

1. Old, W. M. *et al. Mol. Cell* **34,** 115–131 (2009).
2. Gray-Schopfer, V., Wellbrock, C. & Marais, R. *Nature* **445,** 851–857 (2007).
3. Hanahan, D. & Weinberg, R. A. *Cell* **100,** 57–70 (2000).
4. Sarno, S. & Pinna, L. A. *Mol. Biosyst.* **4,** 889–894 (2008).
5. Ritt, D. A. *et al. Curr. Biol.* **17,** 179–184 (2007).
6. Garnett, M. J., Rana, S., Paterson, H., Barford, D. & Marais, R. *Mol. Cell* **20,** 963–969 (2005).
7. Dumaz, N. *et al. Cancer Res.* **66,** 9483–9491 (2006).
8. Specht, K. M. & Shokat, K. M. *Curr. Opin. Cell Biol.* **14,** 155–159 (2002).

PLANT BIOTECHNOLOGY

Zinc fingers on target

Matthew H. Porteus

The existing methods of creating genetically modified plants are inefficient and imprecise. Zinc-finger technology offers the prospect of opening up a swifter and more exact route for crop improvement.

Go into any supermarket in the developed world, and you would be sure that we live in a world of plenty. Yet elsewhere on the globe tens of thousands of people die every day from starvation, and today, as much as ever, ways of improving food production are needed. The ability to create precise genetically modified plants is one approach to that end, and papers by Shukla *et al.*[1] and Townsend *et al.*[2] (pages 437 and 442 of this issue) provide a promising way forward. Both groups exploit the potential of enzymes called zinc-finger nucleases.

Zinc-finger nucleases (ZFNs) are engineered proteins that can be designed to make a single break at a specific site in double-stranded genomic DNA[3]. By hijacking a cell's own repair machinery to repair the break, one can then create specific gene modifications. ZFNs have been used to make site-specific genome modifications in animal cells in three general ways (Fig. 1), and they have also previously been used in research on plants[4–7]. But the new papers[1,2] provide the first demonstration that ZFNs can be used to create plants that have precise changes in endogenous genes and that breed true. This precision is achieved by exploiting the process of homologous recombination, a type of genetic recombination between DNAs of the same or very similar sequence that is naturally used to repair DNA damage and to create genetic diversity.

Shukla *et al.*[1] and Townsend *et al.*[2] respectively

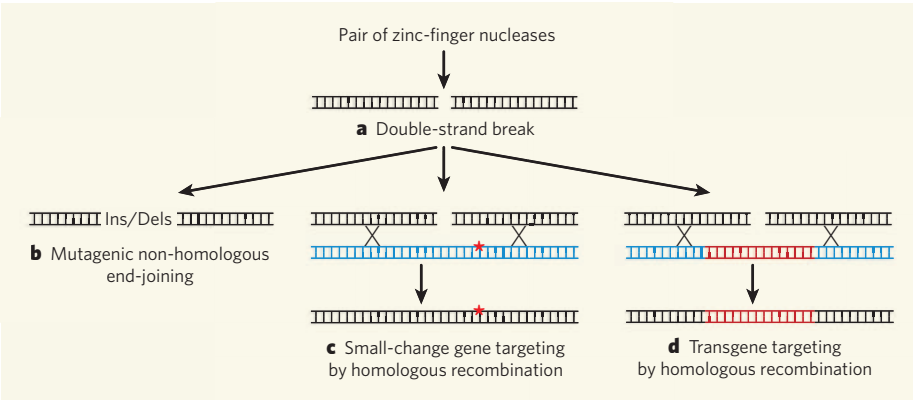


Figure 1 | Site-specific endogenous gene modification with zinc-finger nucleases (ZFNs). **a**, A pair of ZFNs is engineered to bind to a particular genomic target site and create a double-strand DNA break. That break can then be processed in three ways to generate a specific modification of the target gene. **b**, 'Mutagenic non-homologous end-joining' creates small insertions or deletions (Ins/Dels) at the site of the break. **c**, 'Small-change gene targeting' involves the use of homologous recombination to introduce small, defined changes (red star) into the target gene. **d**, 'Transgene targeting' exploits homologous recombination to introduce a different gene into the target genomic region, thereby inactivating the target gene and adding a new trait.

describe modification of the maize gene *IPK1* and the tobacco gene *SuRB*; the various stages are shown in Figure 2. Both groups created a large number of ZFNs and screened them for their ability to recognize the target gene using a series of assays to identify the best pairs of enzymes (a pair being required to make a double-strand break). They then used those ZFNs to create insertions or deletions in the target genes by a procedure called mutagenic non-homologous end-joining; this generates desired mutants at high enough rates that mean that mutants can be readily identified by screening.

Both groups then further increased the specificity of target-gene modification by ZFNs by exploiting the cell's homologous recombination machinery to create precise sequence changes. In this way, Shukla *et al.*[1] introduced a herbicide-resistance gene into the *IPK1* locus. This manipulation had the double benefit of simultaneously inactivating the *IPK1* gene (which means that less of the undesirable end-product, phytate, is produced in the seeds) and creating herbicide tolerance, a process termed 'trait stacking'. They found that 3–20% of the resistant samples had *IPK1*-targeted integrations. The rate increased to 17–100% when they used a targeting strategy that reduced the background of non-targeted events. Importantly, Shukla and colleagues created genetically modified plants that breed true, passing the targeted gene modification to their progeny at Mendelian frequencies.

Townsend *et al.*[2] used homologous recombination to target small changes to the *SuRB* gene. When the changes were 188 base pairs from the ZFN site, they achieved targeting rates of 4.0%. This rate decreased to 0.2% when the targeted changes were 1,541 base pairs from the ZFN-induced double-strand break, consistent with a decrease in targeting effectiveness as the distance from the site of the break increases[8]. Distance from the break might also explain the difference between targeting rates observed in maize and tobacco, as the targeting in maize was made closer to the break (although other factors probably also contribute).

The application of ZFNs to create genetically modified plants is a significant advance. Existing techniques — which variously involve transferred (T) DNA, transposons or chemical mutagenesis — create random mutations, with the subsequent requirement for molecular or phenotypic screening for the desired characteristics. With ZFNs one can create small insertions or deletions at a specific genomic location, insert a new gene (which permits trait stacking, a process that is much more difficult using existing approaches), or create precise point mutations. The time and effort involved in screening is reduced by at least an order of magnitude compared with the other methods.

What are the options for obtaining active ZFNs, the essential elements in this approach? The technically simplest is to negotiate a collaborative agreement with Sangamo/Sigma–Aldrich/Dow Agrosciences to have them make customized ZFNs (the work of Shukla *et al.*[1] is a collaboration between Sangamo and Dow). Whereas ZFNs can be simply purchased from Sigma–Aldrich for research purposes in animals, current licensing agreements between the three companies preclude that option for use in plants. The next simplest option is to adopt the approach of modular assembly[9]. Any lab competent in general molecular-biology techniques can assemble new ZFNs in this way, although it has the drawbacks of relative inefficiency in making active ZFNs and lower specificity of those ZFNs that are active[10,11]. The final option is to use a selection-based approach[12,13], one of which is the OPEN system[6,11] (an open-source initiative, to which both Townsend *et al.*[2] and myself are contributors). The selection method is technically the most demanding of the three strategies, and its track record is less established than that of the Sangamo/Sigma–Aldrich technology.

Although each strategy can generate ZFNs to many target sites, none has yet advanced to the stage where active ZFNs can be generated for every possible site. This limitation, which is the subject of ongoing research, is alleviated somewhat by the fact that successful targeting can occur at a distance from the intended ZFN site. One of the other important aspects of ZFNs is their potential off-target effects. The *SuRB*-directed ZFNs, for example, also created mutations in the homologous *SuRA* gene; and the possibility of low-frequency targeting of *IPK2*, an *IPK1* homologue, could not be ruled out. Combining ZFNs with homologous recombination reduces off-target gene modifications, although this issue is less serious in plants than, for example, in potential ZFN use for gene therapy, because unwanted effects can be bred away.

As well as providing us with food and other materials, plants are increasingly seen as a renewable energy source. These two reports[1,2] provide solid, proof-of-principle evidence that application of ZFNs can improve both understanding of plant biology and efficiency in agriculture. ■

Matthew H. Porteus is in the Departments of Pediatrics and Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.
e-mail: matthew.porteus@utsouthwestern.edu

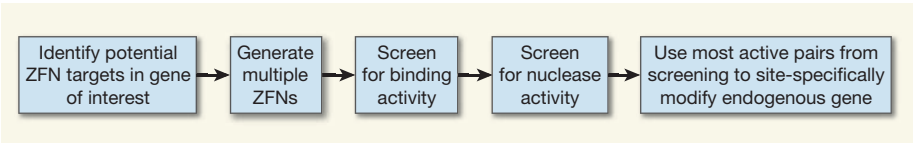


Figure 2 | Use of ZFNs to modify plant genes. This flow chart shows the sequence of steps undertaken by Shukla *et al.*[1] and Townsend *et al.*[2] respectively to target sites in the maize and tobacco genomes. In the final step, site-specific gene modification, Shukla *et al.* used transgene targeting and Townsend *et al.* used small-change gene targeting (see Fig. 1).

1. Shukla, V. K. *et al. Nature* **459,** 437–441 (2009).
2. Townsend, J. A. *et al. Nature* **459,** 442–445 (2009).
3. Porteus, M. H. & Carroll, D. *Nature Biotechnol.* **23,** 967–973 (2005).
4. Lloyd, A., Plaisier, C. L., Carroll, D. & Drews, G. N. *Proc. Natl Acad. Sci. USA* **102,** 2232–2237 (2005).
5. Cai, C. Q. *et al. Plant Mol. Biol.* **69,** 699–709 (2008).
6. Maeder, M. L. *et al. Mol. Cell* **31,** 294–301 (2008).
7. Wright, D. A. *et al. Plant J.* **44,** 693–705 (2005).
8. Elliot, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. *Mol. Cell. Biol.* **18,** 93–101 (1998).
9. Carroll, D., Morton, J. J., Beumer, K. J. & Segal, D. J. *Nature Protoc.* **1,** 1329–1341 (2006).
10. Ramirez, C. L. *et al. Nature Meth.* **5,** 374–375 (2008).
11. Pruett-Miller, S. M., Connelly, J. P., Maeder, M. L., Joung, J. K. & Porteus, M. H. *Mol. Ther.* **16,** 707–717 (2008).
12. Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D. & Wolfe, S. A. *Nature Biotechnol.* **26,** 695–701 (2008).
13. Durai, S., Bosley, A., Abulencia, A. B., Chandrasegaran, S. & Ostermeier, M. *Comb. Chem. High Throughput Screen* **9,** 301–311 (2006).

Q&A

KELLY-MOONEY PHOTOGRAPHY/CORBIS

OBESITY

Causes and control of excess body fat

Jeffrey M. Friedman

Obesity is a major health problem in developed countries and a growing one in the developing world. It increases the risk of diabetes, heart disease, fatty liver and some forms of cancer. A better understanding of the biological basis of obesity should aid its prevention and treatment.

How is obesity diagnosed?

Obesity is defined as excessive adiposity (body fat). Because the historical method for estimating adiposity — calculating a person's buoyancy by measuring their body weight under water — is cumbersome, a surrogate measure known as the body mass index (BMI) is now routinely used. Although BMI (weight in kilograms per square of height in metres) is a convenient measure and useful for assessing the weight status of a population over time, it is often unreliable for assessing an individual's status. This is because it does not distinguish between fat and muscle mass, and the use of height squared, which, like weight, is subject to demographic shifts, is entirely empirical. Recently developed methods for directly assessing body fat, such as air displacement to calculate density, are more reliable and so should replace BMI measurements.

How big a problem is obesity?

It is a global problem. Among Caucasians, the risk of obesity-associated medical complications first becomes evident from actuarial, or mortality, tables in people with a BMI of 25 (overweight), and rises drastically in those with BMIs of 30 and above (obese). In the United States, around one-third of the population has a BMI above 30, and half have a BMI of more than 25. The aggregate economic cost of obesity in this one country is estimated to be in excess of US$60 billion per year, with a large proportion of it attributable to obesity-associated type 2 diabetes. The severity of obesity often increases dramatically when calories become freely available to populations. Hence, obesity is a growing problem in China and India, among other countries, and is likely to become a bigger problem as more nations get richer. In Asian populations, the risk of diabetes and other metabolic diseases often develops at lower BMIs than among Caucasians, amplifying the health consequences of obesity in Asia.

So is there an epidemic of obesity?

No: an epidemic typically denotes a disease that spreads from person to person and has a rapidly increasing incidence. Many trumpet a dramatic increase in the incidence of obesity, but whether this view is true depends on how one looks at the data. Variation in body weight is a continuous trait, whereas obesity is a dichotomous trait. Giving a fixed threshold to a continuous trait — for instance, that a BMI of 25 or above indicates overweight — means that a small shift in the trait's mean value leads to a disproportionate increase in the number of people who exceed the threshold. For example, there were reports in the 1990s of a 33% increase in the incidence of obesity in the United States during the previous decade, strongly supporting the role of lifestyle. What remained unreported was an average weight gain of only 3–5 kilograms over the same decade in the whole population. So the secular trend towards obesity is less profound than is generally appreciated.

Are you saying that obesity is not a disease of lifestyle?

Lifestyle or environment is probably a necessary but insufficient factor in obesity. Going back to the earlier US example, although the vast majority of individuals there have unlimited access to calories, only half of the population is overweight or obese. A key question therefore is: when provided with unlimited calories, why do only some people consume more than others, becoming obese?

And the answer is?

Although many believe that food intake is primarily a voluntary, conscious behaviour, evidence suggests that the balance between energy intake and output is largely controlled by a powerful, unconscious biological system. It stands to reason that a biological system that maintains energy balance would be under evolutionary pressure, making it weigh up the relative risks and benefits of different amounts of fat. In a hunter–gatherer society, too little fat would put an individual at risk of starvation, whereas too much of it would increase the risk of both predation and serious disease. Thus, genetic variants that contribute to leanness or obesity could both be beneficial to a population, depending on the environmental conditions. This might explain why populations that have been historically undernourished often become the most obese when suddenly provided with unlimited calories.

What about consciously balancing food intake and energy expenditure?

On the basis of the laws of thermodynamics, body weight could be controlled in this way. But the biological system that balances adipose-tissue mass resists weight change in either direction, partly by regulating the unconscious drive to eat. In the short term, therefore, a motivated individual will lose weight by reducing food intake and/or increasing energy expenditure. Eventually, however, biological factors supervene and confer a powerful, unconscious impulse to eat more until the individual returns to his or her starting weight. This is analogous to consciously holding one's breath; inevitably, the basic drive to breathe dominates the conscious motivation. Consider variations in weight among groups of individuals over time — say a year, during which an individual would consume roughly one million calories. Weight remains remarkably stable, far exceeding an individual's conscious ability to monitor their food intake and energy expenditure. A challenging question is how neural circuits that underlie the basic drive to eat interact with those that represent the conscious wish to alter one's weight.

What is the evidence for a biological basis for obesity?

Classically, a genetic contribution to a human trait is quantified by comparing the trait's variation between identical and non-identical twins. Using this approach, the heritability of obesity — percentage of variation due to genetic factors — ranges between 70% and 80%. These values exceed those for most other traits that are commonly accepted to have a biological basis, including diabetes, heart disease and cancer. Indeed, the only trait with consistently higher heritability than obesity is height. Adoption studies also support the contribution of genes to obesity: adopted children's weight more closely resembles that of their biological, rather than their adoptive, parents. Nonetheless, no study can completely attribute obesity to genes. Because no one becomes obese if they are starved, the environment — primarily, free access to calories — is probably a permissive factor that sets the stage for the genetically predisposed to become obese.

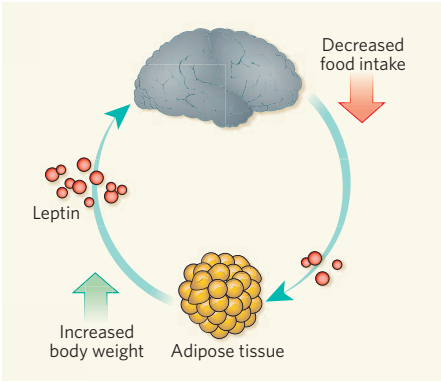


Figure 1 | Leptin and the control of body fat. With increased body weight, adipose tissue secretes higher levels of leptin. This hormone then travels to the brain, where it binds to leptin receptors in various regions, including the hypothalamus. The result is a sensation of satiety and so a decrease in food intake. Conversely a reduction in body weight lowers leptin levels and increases food intake. Thus, relative constancy of weight can be maintained.

What genes have been implicated?

Several genes, when mutated, cause obesity in humans and animals. These genes are generally components of the system that regulates energy balance. For instance, the *ob* gene encodes leptin — a hormone made in adipose tissue that acts on many physiological systems, including brain centres that control food intake and energy expenditure. With an increase or decrease in body fat, leptin levels fluctuate accordingly, leading to a respective reduction or increase in food intake (Fig. 1). Mice with mutations in *ob* fail to produce leptin and show a threefold increase in weight and a fivefold increase in body fat compared with normal mice. Humans with mutations in this gene, or in the gene for the leptin receptor, can also become massively obese. The leptin receptor is located in the hypothalamus and elsewhere in the brain, as well as in some peripheral tissues. Injury to the hypothalamus can cause obesity, partly by destroying neurons that express the leptin receptor. A class of leptin-activated neurons in the hypothalamus express the neuropeptide precursor POMC, mutations in which, or in its receptor MC4, also cause obesity.

Do only single-gene mutations cause obesity?

No. Some 5–10% of morbid obesity (BMI of 40 or more) is due to defects in the above genes and in other genes that function in the brain circuits, including that encoding the neuropeptide BDNF. This is an unusually high frequency for the Mendelian inheritance of a complex trait. In the rest of the population, however, a combination of genes and their interaction with the environment is thought to cause weight variation. Several genes that affect weight have been identified through genome-wide association studies. One such gene is *FTO*, the function of which is unknown, but DNA-sequence variations in it can account for a 3–5-kilogram weight difference. The fraction of genes such as *FTO* that have been shown to contribute to obesity in the general population is relatively low. So it remains unclear whether such genes contribute to obesity through many different single-gene mutations or through potentially complex interactions involving several genes, each having small effects on their own (or a combination of both).

Is there a difference in the metabolism of lean and obese people?

Not if one measures only lean body mass. But when obese people lose weight, their energy expenditure is reduced disproportionately to their change in weight. These 'reduced obese' people use less energy than lean individuals of that weight who have not been obese. And to maintain their reduced weight, they must consume fewer calories than their initially lean counterparts. This disadvantage in itself undoubtedly contributes to the high rate of recidivism after dieting, especially as it occurs at a time when the basic drive to eat is activated by reduced leptin levels.

What hormones, other than leptin, are involved?

Whereas leptin maintains constant energy stores over long periods, there's another system that maintains relatively constant levels of nutrients in the blood in the short term, for instance throughout the day. This system, which controls both hunger and satiety, consists of many blood-borne and neural signals. The blood-borne signals include metabolites — such as glucose, and possibly amino acids and fatty acids — and hormones of the digestive system, including the stomach hormone ghrelin and intestinal peptides such as GLP-1, peptide YY, cholecystokinin, bombesin and amylin. These short-term signals act on neurons in the brainstem and hypothalamus to regulate both food intake and the intervals between meals. The short- and long-term systems interact extensively.

So it seems that several tissues and organs regulate weight?

Indeed. Food intake and body weight are controlled by an intercalated feedback loop. Signals from numerous tissues that together form the short- and long-term systems — including adipose tissue and the gut — travel to integratory brain centres, where they are decoded. The neural pathways then control food intake and metabolism in several

peripheral tissues (Fig. 2). Although substantial progress has been made in defining the neural pathways that control food intake, less is known about the circuits that regulate energy expenditure, and fat and glucose metabolism.

If predisposition to obesity is determined by our genes and metabolism, why diet or exercise?

Weight loss alleviates obesity-associated medical complications, with even modest losses of 5–7 kilograms having disproportionate benefits to health. Many people can achieve this amount of weight loss, partly because the potency of the biological factors that resist changes in weight is greatest after larger amounts have been lost. So my advice to obese individuals is the same as I would offer anyone: do what you can to improve your health. Eat a heart-healthy diet, begin a programme of physical activity, and try to lose as much weight as is required to improve your health, without feeling compelled to 'normalize' your weight.

Can gut microorganisms cause obesity?

Some studies have suggested a small but significant contribution of the gut microbiota, although the underlying mechanism is unknown. The contribution of such organisms, however, seems to be much smaller than that of the host genes. In animals, certain viruses can cause obesity by damaging the hypothalamus, but this effect has not been seen in humans.

Does the cause of obesity affect its severity?

Demonstrably so among those with specific obesity-related mutations. Patients with mutations in leptin or in the leptin receptor, for example, are more obese than those with mutations in POMC, MC4 or BDNF. Among obese people in the general population, there is probably a similar or even greater degree of variation. To identify such differences, respective genetic determinants in subgroups of obese people must be identified — an endeavour that is already under way.

What anti-obesity therapies are out there, and how effective are they?

One effective therapy is bariatric surgery to modify the anatomy of the gastrointestinal tract, thereby reducing food intake and/or absorption. Because all bariatric procedures can potentially cause serious morbidity and even death, this treatment is typically reserved for those with severe medical problems. Besides, even after surgery, most patients remain clinically obese (BMI more than 30), despite the marked reduction in their food intake. This fact highlights the biological difference between the morbidly obese and individuals of average weight. The reason for the effectiveness of bariatric surgery is unclear.

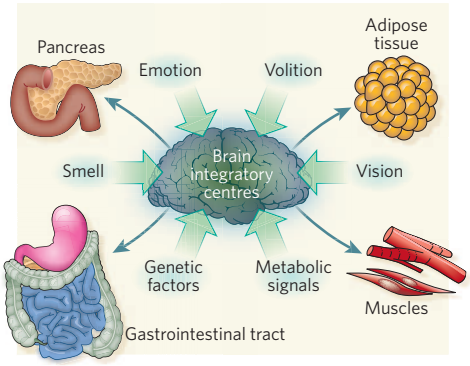


Figure 2 | Feeding is a complex, motivational behaviour. Several behavioural, genetic and metabolic signals regulate feeding through signals that travel from distinct peripheral tissues to integratory centres in the brain. After processing these signals, the brain modifies feeding and also sends appropriate commands to specific peripheral tissues to regulate metabolism.

Many believe that its benefits are primarily due to alterations in neural, metabolic and hormonal signals from the gastrointestinal tract to the brain, rather than a mechanical alteration that physically limits food intake. A major goal is to identify the specific signals that are altered by this procedure. The alternative procedure of liposuction is effective only in the short term, as the lost fat is eventually regained, probably because leptin levels are lowered after the removal of large amounts of fat, and so food intake increases. Leptin-replacement therapy in those deficient in this hormone is another approach that results in dramatic weight loss in animals and in the small number of humans with leptin mutations; to date, there are no known disadvantages of this treatment.

What are the hottest developments in obesity research?

For one, our increasing ability to identify specific neural pathways that control feeding and to assess how modulating the activity of such neurons affects feeding behaviour. For instance, manipulation with light, involving the light-activated ion channel known as channel rhodopsin, can be used to examine the effect of activating or inhibiting a neuron. Also, functional imaging has recently been used to map the human brain regions that control eating. Such studies have revealed leptin-mediated changes in neural circuits that control reward-associated behaviour, providing a link between neurobiology and psychology. As for research into anti-obesity therapy, emerging data show that leptin and amylin together produce a potent signal to induce substantial weight loss. The aim now is to evaluate the safety and long-term efficacy of this combined therapy.

And what are the most pertinent remaining questions?

A crucial objective is to understand the way in which diverse inputs lead to a single behavioural response — feeding. We do not know how this complex information is represented in the brain centres that control eating, or even where exactly these centres are. Answers to these questions may eventually reveal how and why, at a neurobiological level, the conscious desire to lose weight is so often dominated by the basic drive to eat. Another outstanding goal is to identify all of the genetic variants that contribute to differences in weight.

What might the future hold?

When there is no appreciable risk of starvation, obesity simply leads to disease, and so evolution should select against it. Indeed, on the Pacific island of Nauru, a profound increase in the incidence of diabetes after the introduction of a high-calorie diet was followed by a decrease, suggesting that there was evolutionary selection against diabetes when the incidence became very high. There is also evidence from the United States that the incidence of obesity may be reaching a plateau. I therefore anticipate that body weight will stabilize in the population over the coming decades.

So is taking action justified?

Regardless of future trends, developing effective and safe anti-obesity therapies is essential and feasible. I consider it less likely that drugs to normalize the weight of morbidly obese people will be developed any time soon; and anyway, whether such drugs would bring added health benefits is unclear. The focus should be on designing treatments that can stably maintain moderate weight loss, improving an individual's health. With the growing realization that mainly biological factors contribute to obesity, it is hoped that this condition will be de-stigmatized, reducing the compulsion by obese individuals to achieve an (arbitrary) ideal, lean weight and instead motivating them to focus on improving their health. This outcome will best serve our larger interests and reflect better on all of us. ■

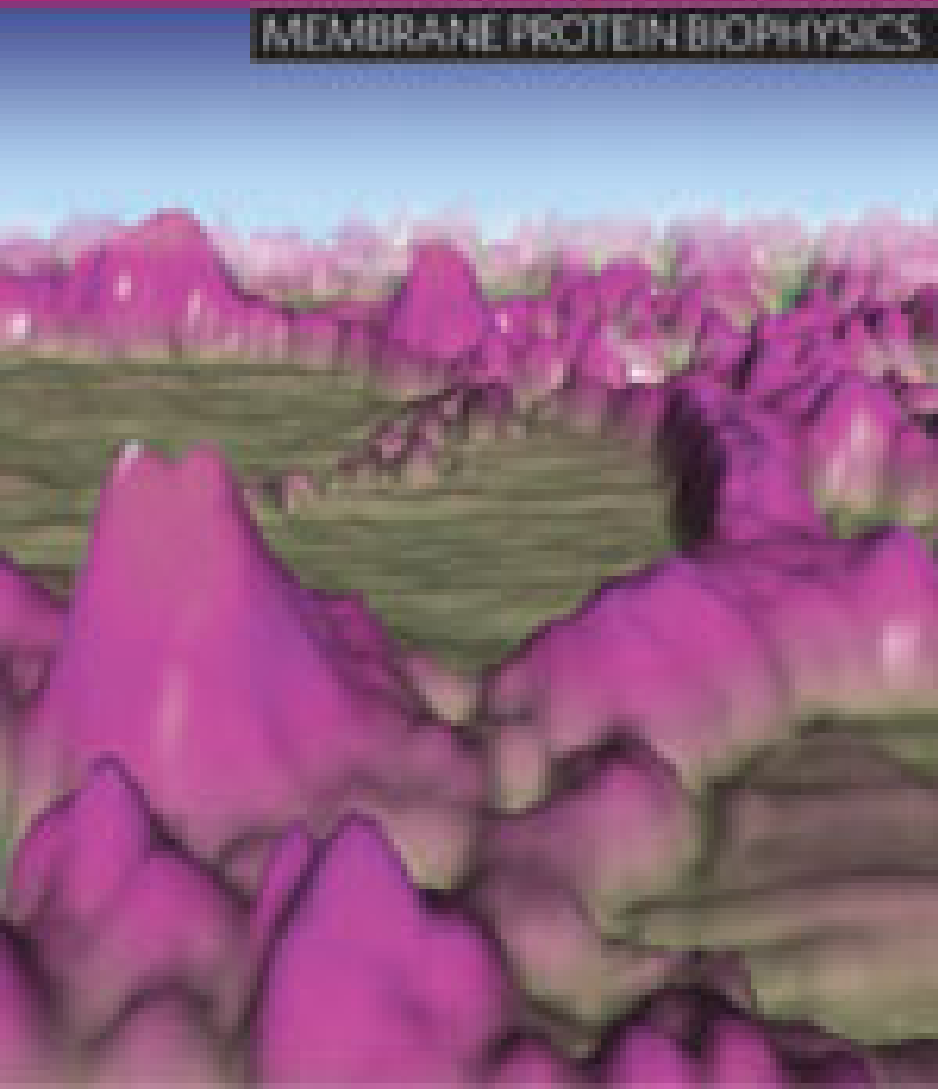
Jeffrey M. Friedman is at the Howard Hughes Medical Institute, The Rockefeller University, New York, New York 10065, USA.
e-mail: friedj@rockefeller.edu

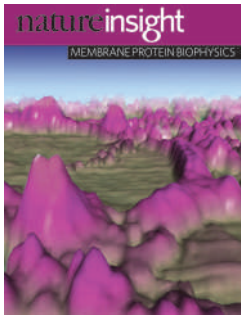
The author declares competing financial interests. See online article for details.

FURTHER READING
Friedman, J. M. Modern science versus the stigma of obesity. *Nature Med.* **10,** 563–569 (2004).
Eckel, R. H. Clinical practice. Nonsurgical management of obesity in adults. *N. Engl. J. Med.* **358,** 1941–1950 (2008).
Coll, A. P., Farooqi, I. S. & O'Rahilly, S. The hormonal control of food intake. *Cell* **129,** 251–262 (2007).
Speakman, J. R. A nonadaptive scenario explaining the genetic predisposition to obesity: the 'predation release' hypothesis. *Cell Metab.* **6,** 5–12 (2007).
Farooqi, S. & O'Rahilly, S. Genetics of obesity in humans. *Endocr. Rev.* **27,** 710–718 (2006).
Friedman, J. M. A war on obesity, not the obese. *Science* **299,** 856–858 (2003).

nature insight

MEMBRANE PROTEIN BIOPHYSICS





Cover illustration
Membrane proteins (peaks) embedded in a lipid bilayer. (Courtesy of H. Schillers and H. Oberleithner, University Hospital of Muenster/SPL; Artwork by N. Spencer).

Editor, *Nature*
Philip Campbell

Publishing
Nick Campbell
Claudia Banks

Insights Editor
Lesley Anson

Production Editor
Davina Dadley-Moore
Anna York

Senior Art Editor
Martin Harrison

Art Editor
Nik Spencer

Sponsorship
Amélie Pequignot
Reya Silao

Production
Jocelyn Hilton

Marketing
Elena Woodstock
Emily Elkins

Editorial Assistant
Emma Gibson

MEMBRANE PROTEIN BIOPHYSICS

Cells and organelles are enclosed by lipid bilayers, which separate them from their external environments. Embedded in these membranes are specialized proteins that span the full width of the bilayer and facilitate communication between the two sides.

These membrane proteins carry out a multitude of tasks, including signal transduction, transportation of small molecules and catalysis of enzymatic reactions. Their architectural repertoire is equally diverse, ranging from a single transmembrane-spanning domain in receptor tyrosine kinases, to the multisubunit ATP synthase.

Although membrane proteins constitute more than a quarter of all known proteins, their dependence on lipids renders them difficult to crystallize, and biophysicists have traditionally relied on indirect techniques to investigate their structural framework.

But in 1985, Johann Deisenhofer and his colleagues succeeded in solving the first atomic-resolution structure of a membrane protein — a photosynthetic reaction centre. Since then, almost 200 unique membrane-protein structures have been revealed.

This *Nature* Insight presents a snapshot of the current position of the field by focusing on a varied selection of transporters, transducers and enzymes of known structure. The Reviews discuss the intriguing mechanisms by which these molecular machines function, as well as the emerging role of lipids in these processes.

We thank all the authors for contributing their visions for the field and look forward to the multitude of molecular mechanisms that remain to be uncovered.

Lesley Anson, Senior Editor

OVERVIEW

REVIEWS

nature insight

NATURE|Vol 459|21 May 2009|doi:10.1038/nature08142

Biophysical dissection of membrane proteins

Stephen H. White[1]

The first atomic-resolution structure of a membrane protein was solved in 1985. Twenty-four years and more than 180 unique structures later, what have we have learned? An examination of the atomic details of several diverse membrane proteins reveals some remarkable biophysical features and suggests that we can expect to achieve much more in the decades to come.

Two events define the beginning of the modern era of membrane-protein biophysics: the determination of the three-dimensional structure of bacteriorhodopsin[1] at low resolution by Richard Henderson and Nigel Unwin in 1975, and the atomic-resolution structure of the *Rhodopseudomonas viridis* photosynthetic reaction centre[2] by Johann Deisenhofer and Hartmut Michel in 1985. Although spectroscopic studies[3,4] of isolated membranes had suggested in the mid-1960s that the proteins of plasma membranes are largely α-helical, Henderson and Unwin's 1975 structure[1] established unequivocally the existence of transmembrane α-helices in a membrane protein. Deisenhofer and Michel's structure confirmed the presence of transmembrane α-helices and allowed an atomic-level interpretation of biophysical data for the first time. As important as the structure was on its own, the more important accomplishment may simply have been to demonstrate that membrane proteins could be crystallized, thereby opening the way for atomic-resolution X-ray crystallography.

As other groups picked up the challenge, progress accelerated (Fig. 1a) such that now, 24 years later, we have high-resolution structures for more than 180 unique proteins (Fig. 1b). What have we learned about membrane-protein biophysics as a result? In this Overview, I extract some general features of membrane-protein biophysics from the five reviews that follow in this Insight, which dissect the operation of several membrane proteins at the atomic level. The clearest feature is that interactions with the lipid bilayer are important, but no general principles about these have yet emerged. Symmetry, which arises from gene duplications, is seen to be a building block of transporter function. In addition, we find that water can penetrate deeply into membrane proteins, despite the non-polar character of the bilayer. As expected from their biosynthetic pathways, all plasma-membrane proteins are α-helical bundles, but there are wide variations in geometry, related in part to water penetration and interactions with the bilayer. Seven-helix bundles, which are hallmarks of G-protein-coupled receptors, seem ideally minimalist and are extremely versatile signalling platforms.

Transporter structural motifs

Sodium-coupled transporters, which are reviewed by Eric Gouaux (see page 347), are an important class of membrane protein, exhibiting pseudo-two-fold symmetry and deep water penetration. These so-called secondary transporters are 'couplers' because they couple the energetic 'uphill' movement of one solute to the 'downhill' movement of another solute. For sodium-coupled transporters, the energy gained from the movement of sodium ions down an electrochemical gradient is used to concentrate substrates, such as aspartate[5] and leucine[6], on one side of the membrane. These transporters, which take advantage of the two-fold structural symmetry, work as 'rocker switches' to provide alternating access to the two sides of the membrane[7]. As Gouaux points out, a broad range of transporters share these features despite having low sequence identity and differing numbers of transmembrane helices. For example, the vSGLT sodium/galactose symporter[8] has 14 transmembrane helices and the Mhp1 benzyl-hydantoin transporter[9] has 12 transmembrane helices, but the transporter function is carried out by pseudo-two-fold-related 5+5 transmembrane repeats.

Signalling and seven-helix bundles

Another class of couplers are the G-protein-coupled receptors (GPCRs). These seven-helix membrane proteins receive an optical or chemical signal on the extracellular membrane surface and initiate G-protein-based signalling cascades in the cytoplasm. GPCRs, which are categorized into six subclasses, form the largest single class of eukaryotic membrane proteins and are the targets of many of the drugs being developed by pharmaceutical companies[10]. Because of their pharmaceutical importance, there have been extensive efforts — dating back to 1992 at least[11] — to build molecular models of ligand-activated GPCRs to guide drug discovery. The first models were based on bacteriorhodopsin, which has seven transmembrane helices and a covalently linked retinal whose photoexcitation drives proton transport. But when the structure of bovine rhodopsin, a true GPCR, was determined[12], modellers turned their attentions to that, with little success[13]. Finally, in 2007, the first structures of ligand-activated GPCRs became available, starting with human β_2 adrenergic receptor co-crystallized with Fab5 antibodies[14]. On page 356, Daniel Rosenbaum, Søren Rasmussen and Brian Kobilka review the structural and mechanistic insights into GPCR function gained from these structures, which include higher-resolution structures of β_2 adrenergic receptor[15,16], turkey β_1 adrenergic receptor[17], human A_{2A} adenosine receptor[18], opsin[19] and squid rhodopsin[20].

A recent mini-review provides a close structural comparison of GPCR structures[21]. The lesson that emerges is that the seven-transmembrane-helix bundle is extremely adaptable, and perhaps also minimalist. Why minimalist? Bacteriorhodopsin can be refolded in lipid vesicles from separated helices (A, B and CDEFG)[22]. Using retinal absorbance as a measure of folding, neither A+CDEFG nor B+CDEFG gives native spectra, whereas A+B+CDEFG does. This implies that seven is the minimum number of helices needed to maintain the native environment of retinal. It may be that seven helices provide ample space for a wide range of ligands through relatively minor helix distortions and reorientations, without the need to increase or decrease the number of transmembrane helices.

[1]Department of Physiology and Biophysics, and Center for Biomembrane Systems, University of California, Irvine, California 92697, USA.

Helix hairpins

A physicist friend once opined that the concept sketches, such as rocker switches, drawn by biologists must be hopelessly naive. Surely the real mechanisms must be far more subtle, drawing on sophisticated physics not yet revealed? Well, membrane protein machines really do function largely according to the principles embodied in the sketches (which, of course, summarize years of painstaking work, a fact overlooked by my friend). This is certainly true of F_OF_1 ATPase (or simply ATP synthase). Paul Boyer introduced in the 1980s the concept of rotary catalysis[23] by F_1 ATPase; John Walker and colleagues[24] provided the nearly complete structural details of rotary catalysis by solving the structure of the ATP synthase at atomic resolution.

ATP synthase generates ATP from proton electrochemical gradients in mitochondria. The ATP-producing F_1 domain is a rotor with a shaft that passes through the membrane-embedded F_O sector, which is composed of ten transmembrane two-helix c-subunits. The multiple two-helix-hairpin motif is apparently ideal for flexibly stabilizing large complexes with circular symmetry, such as light-harvesting complexes[25,26]. ATP synthases of different types can have up to 14 c-subunit hairpins[27]. Although the exact biochemical reason for variations in c-subunit number is not understood, structurally it seems to be an efficient way to expand a boundary of a membrane protein. Need a larger boundary? Just add a few more hairpins. Protons that pass through F_O on their way down the electrochemical gradient drive the F_1 rotor to produce ATP via the mechanical energy produced. Remarkably, the rotation of the rotor can be observed and quantified by attaching a fluorescent actin filament to the γ-subunit shaft[28]. Wolfgang Junge and his colleagues Hendrik Sielaff and Siegfried Engelbrecht describe the structure and operation of this remarkable machine in their review (see page 364). Junge provides excellent movies of the synthase in action as Supplementary Information. More movies and structural information can be found at www.biologie.uni-osnabrueck.de/biophysik/junge/picsmovs.html and www.mrc-dunn.cam.ac.uk/research/atp_synthase.

Bilayer distortion

Intramembrane-cleaving proteases (iCLIPs), first identified by Michael Brown, Joseph Goldstein and colleagues[29,30], irreversibly cleave transmembrane helices to release tethered signalling domains intra- or extracellularly. For example, site-2 protease (S2P) cleaves the cytoplasmic domain of the mammalian sterol regulatory element-binding protein (SREBP). The liberated domain then travels to the nucleus to activate genes involved in cholesterol and fatty-acid biosynthesis. An iCLIP of current interest is the catalytic subunit of γ-secretase, presenilin, which liberates the amyloid-β peptide. S2P and presenilin are examples of metallo- and aspartyl proteases, respectively. A third class of iCLIP comprises the rhomboid serine proteases, which in *Drosophila* liberate the epidermal growth factor domain by cleaving the single-transmembrane Spitz protein.

The conceptual difficulty with iCLIPs is that they must carry out hydrolytic reactions within the hydrophobic membrane interior. The mystery of how they achieve this is emerging from the recently published structures of S2P[31] and rhomboid proteases from the bacteria *Escherichia coli*[32–34] and *Haemophilus influenzae*[35], as discussed by Elinor Erez, Deborah Fass and Eitan Bibi in their comprehensive but succinct review of intramembrane proteases (see page 371). As with sodium-coupled transporters, there is deep penetration of water into the heart of the rhomboid, allowing peptide-bond hydrolysis by a catalytic dyad composed of serine and histidine. An unusual (perhaps even unique) feature of the six-helix rhomboids is the rather large L1 helical loop between helix 1 and helix 2 that protrudes parallel to the membrane plane with deep penetration into the bilayer interface. The exact function of this loop is uncertain, but it may be involved in the regulation of proteolysis. Helix 5 sits at the membrane entrance to the catalytic site, but is on the opposite side of the protein to L1. Molecular dynamic simulations[36] suggest that there is dynamic coupling between L1 and helix 5, consistent with a regulatory role for L1. The simulations also show that the irregular shape of the protein causes significant bilayer deformation in the vicinity of the active site, and that hydrogen-bond interactions with bilayer lipids affect protein orientation and dynamics. Distortion of the lipids in the vicinity of the active site may be necessary to admit both water and transmembrane-helix substrates.

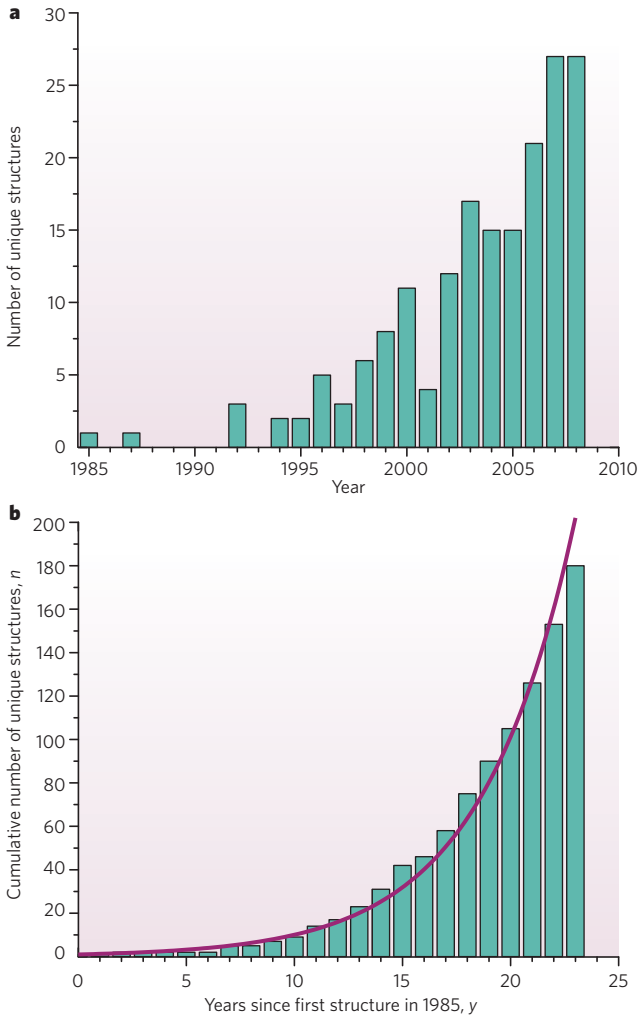


Figure 1 | Progress in determining membrane protein structures. Only unique structures are included in the statistics. Proteins of the same type from different species are included, but structures of mutagenized versions of proteins are excluded, as are proteins that differ only in terms of substrate bound or physiological state. **a**, The number of structures reported each year since 1985. **b**, The bars represent the cumulative number (n) of structures plotted against the number of years (y) since the first structure was reported. The solid curve is the best fit to the equation $n = \exp(ay)$, where $a = 0.23$; the reduced χ^2 of the fit is 0.6. Data are from a curated database of membrane proteins of known structure at http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html.

Restoring the bilayer to membrane proteins

The function of rhomboid proteases requires careful consideration of the effect of the lipid membrane itself, within which all membrane proteins must exist and function. The difficulty with membrane-protein crystallographic structures is that we never see the surrounding membrane, except to the extent that a few lipids or solubilizing detergent molecules may form part of a crystal's unit cell. A powerful approach to understanding a membrane protein in context is to restore the membrane and its lipids computationally, which can be done either by atomic-level molecular dynamics simulations[37] or by continuum mechanics. Rob Phillips, Tristan Ursell, Paul Wiggins and Pierre Sens describe the use of the latter approach (see page 379), which is especially useful for understanding the interplay between tension-gated ion channels and the membrane. It is also useful for describing the

interactions between proteins in the crowded membrane environment[38]. A fundamental question for the future is how to move seamlessly between continuum and atomic-level simulations.

Structural themes

Some unexpected structural themes have emerged from the X-ray structures of membrane proteins discussed here. The bacteriorhodopsin structure, finally solved to atomic resolution by electron crystallography[39] in 1996 and by X-rays[40] in 1997, revealed the transmembrane-helix bundle as the fundamental structural motif of plasma-membrane proteins (bacterial, mitochondrial and chloroplast outer-membrane proteins generally have a β-barrel motif[41,42]). The seven-helix bundle characteristic of bacteriorhodopsin is a remarkably versatile motif, as is apparent from G-protein-coupled receptors. Proteins involved in proton and electron transport form very tight 'waterproof' bundles, whereas transporters such as the *E. coli* lactose permease[43] and the mitochondrial ADP/ATP carrier[44] have large, water-filled cavities that extend almost completely across the membrane. Sodium-coupled transporters share this feature, albeit to a lesser extent. Also discovered in the first structure of an ion channel, the KcsA potassium channel[45], these water-filled cavities complicate the prediction of membrane-protein topology by hydropathy analysis.

Another theme to emerge is an internal structural repeat and inversion of the first half of the protein to form a helix bundle with a pseudo-two-fold symmetry about an axis parallel to the membrane plane. This pattern apparently arises from ancient gene duplications, as described in detail in a recent review[46]. In a few cases, such as the *E. coli* EmrE multidrug transporter[47], the two-fold symmetry may arise from two dual-topology monomers inserted with opposite topologies[48] to form the functional homodimer. Finally, from studies of voltage-sensor domains[49] that control voltage-dependent ion channels and enzymes, and which can act as ion channels themselves, we now know that charged amino acids, especially arginine, can be buried directly in the lipid bilayer as a result of salt bridges and lipid polar-group interactions.

The biophysical motifs discussed here are highlights from a rich menu of structures. The 180 unique structures available at the end of 2008 have revolutionized our understanding of membrane proteins. What does the future hold? What new motifs and folds will appear? James Bowie and colleagues[50] have estimated that about 1,700 membrane-protein structures are needed to account for each structural family. At the present pace and level of technology, that will take about 30 years. So stay tuned for the remarkable progress of the past 24 years to continue. ■

1. Henderson, R. & Unwin, P. N. T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257,** 28–32 (1975).
2. Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3Å resolution. *Nature* **318,** 618–624 (1985).
3. Maddy, A. H. & Malcolm, B. R. Protein conformations in the plasma membrane. *Science* **150,** 1616–1618 (1965).
4. Lenard, J. & Singer, S. J. Protein conformation in cell membrane preparations as studied by optical rotatory dispersion and circular dichroism. *Proc. Natl Acad. Sci. USA* **56,** 1828–1835 (1966).
5. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
6. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporter. *Nature* **437,** 203–205 (2005).
7. Mitchell, P. A general theory of membrane transport from studies of bacteria. *Nature* **180,** 134–136 (1957).
8. Faham, S. *et al.* The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/Sugar symport. *Science* **321,** 810–814 (2008).
9. Weyand, S. *et al.* Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science* **322,** 709–713 (2008).
10. Lagerström, M. C. & Schiöth, H. B. Structural diversity of G-protein-coupled receptors and significance for drug discovery. *Nature Rev. Drug Discov.* **7,** 339–357 (2008).
11. Pardo, L., Ballesteros, J. A., Osman, R. & Weinstein, H. On the use of the transmembrane domain of bacteriorhodopsin as a template for modeling the three-dimensional structure of guanine nucleotide-binding regulatory protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **89,** 4009–4012 (1992).
12. Palczewski, K. *et al.* Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289,** 739–745 (2000).
13. Oliveira, L., Hulsen, D., Hulsik, D. J., Paiva, A. C. M. & Vriend, G. Heavier-than-air flying machines are impossible. *FEBS Lett.* **564,** 269–273 (2004).
14. Rasmussen, S. G. F. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–388 (2007).
15. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
16. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
17. Warne, T. *et al.* Structure of a β-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–492 (2008).
18. Jaakola, V.-P. *et al.* The 2.6 angstrom crystal structure of a human A_{2A} adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).
19. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H.-W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–188 (2008).
20. Murakami, M. & Kouyama, T. Crystal structure of squid rhodopsin. *Nature* **453,** 363–368 (2008).
21. Hanson, M. A. & Stevens, R. C. Discovery of new GPCR biology: One receptor structure at a time. *Structure* **17,** 8–14 (2009).
22. Kahn, T. W. & Engelman, D. M. Bacteriorhodopsin can be refolded from two independently stable transmembrane helices and the complementary five-helix fragment. *Biochemistry* **31,** 6144–6151 (1992).
23. Boyer, P. D. The ATP synthase: A splendid molecular machine. *Annu. Rev. Biochem.* **66,** 717–749 (1997).
24. Stock, D., Leslie, A. G. W. & Walker, J. E. Molecular architecture of the rotary motor in ATP synthase. *Science* **286,** 1700–1705 (1999).
25. Koepke, J., Hu, X. C., Muenke, C., Schulten, K. & Michel, H. The crystal structure of the light-harvesting complex II (B800- 850) from *Rhodospirillum molischianum*. *Structure* **4,** 581–597 (1996).
26. McDermott, G. *et al.* Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* **374,** 517–521 (1995).
27. Weber, J. & Senior, A. E. ATP synthesis driven by proton transport in F_1F_O-ATP synthase. *FEBS Lett.* **545,** 61–70 (2003).
28. Noji, H., Yasuda, R., Yoshida, M. & Kinosita, K. Jr Direct observation of the rotation of F_1-ATPase. *Nature* **386,** 299–302 (1997).
29. Sakai, J. *et al.* Sterol-regulated release of SREBP-2 from cell membranes requires two sequential cleavages, one within a transmembrane segment. *Cell* **85,** 1037–1046 (1996).
30. Rawson, R. B. *et al.* Complementation cloning of S2P, a gene encoding a putative metalloprotease required for intramembrane cleavage of SREBPs. *Mol. Cell* **1,** 47–57 (1997).
31. Feng, L. *et al.* Structure of a site-2 protease family intramembrane metalloprotease. *Science* **318,** 1608–1612 (2007).
32. Wang, Y., Zhang, Y. & Ha, Y. Crystal structure of a rhomboid family intramembrane protease. *Nature* **444,** 179–183 (2006).
33. Wu, Z. *et al.* Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry. *Nature Struct. Mol. Biol.* **13,** 1084–1091 (2006).
34. Ben-Shem, A., Fass, D. & Bibi, E. Structural basis for intramembrane proteolysis by rhomboid serine proteases. *Proc. Natl Acad. Sci. USA* **104,** 462–466 (2007).
35. Lemieux, M. J., Fischer, S. J., Cherney, M. M., Bateman, K. S. & James, M. N. G. The crystal structure of the rhomboid peptidase from *Haemophilus influenzae* provides insight into intramembrane proteolysis. *Proc. Natl Acad. Sci. USA* **104,** 750–754 (2007).
36. Bondar, A.-N., del Val, C. & White, S. H. Rhomboid protease dynamics and lipid interactions. *Structure* **17,** 395–405 (2009).
37. Lindahl, E. & Sansom, M. S. P. Membrane proteins: molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **18,** 425–431 (2008).
38. Engelman, D. M. Membranes are more mosaic than fluid. *Nature* **438,** 578–580 (2005).
39. Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **259,** 393–421 (1996).
40. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipidic cubic phases. *Science* **277,** 1676–1681 (1997).
41. Buchanan, S. K. β-Barrel proteins from bacterial outer membranes: Structure, function and refolding. *Curr. Opin. Struct. Biol.* **9,** 455–461 (1999).
42. Schulz, G. E. β-Barrel membrane proteins. *Curr. Opin. Struct. Biol.* **10,** 443–447 (2000).
43. Abramson, J. *et al.* Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301,** 610–615 (2003).
44. Pebay-Peyroula, E. *et al.* Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature* **426,** 39–44 (2003).
45. Doyle, D. A. *et al.* The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* **280,** 69–77 (1998).
46. von Heijne, G. Membrane-protein topology. *Nature Rev. Mol. Cell Biol.* **7,** 909–918 (2006).
47. Schuldiner, S. EmrE, a model for studying evolution and mechanism of ion-coupled transporters. *Biochim. Biophys. Acta* **1794,** 748–762 (2009).
48. Rapp, M., Seppälä, S., Granseth, E. & von Heijne, G. Emulating membrane protein evolution by rational design. *Science* **315,** 1282–1284 (2007).
49. Swartz, K. J. Sensing voltage across lipid membranes. *Nature* **456,** 891–897 (2008).
50. Oberai, A., Ihm, Y., Kim, S. & Bowie, J. U. A limited universe of membrane protein families and folds. *Protein Sci.* **15,** 1723–1734 (2006).

Acknowledgements This work was supported in part by grants from the National Institute of General Medical Science and the National Institute of Neurological Disorders and Stroke.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (stephen.white@uci.edu).

Unlocking the molecular secrets of sodium-coupled transporters

Harini Krishnamurthy[1], Chayne L. Piscitelli[1,2] & Eric Gouaux[1,3]

Transmembrane sodium-ion gradients provide energy that can be harnessed by 'secondary transporters' to drive the translocation of solute molecules into a cell. Decades of study have shown that such sodium-coupled transporters are involved in many physiological processes, making them targets for the treatment of numerous diseases. Within the past year, crystal structures of several sodium-coupled transporters from different families have been reported, showing a remarkable structural conservation between functionally unrelated transporters. These atomic-resolution structures are revealing the mechanism of the sodium-coupled transport of solutes across cellular membranes.

A multitude of transmembrane transporter proteins have evolved to catalyse the movement of small polar or charged molecules across the hydrophobic barrier of the membrane bilayer[1]. A large class of these proteins, termed secondary transporters, use the discharge of an ionic gradient to power the 'uphill' translocation of solute molecules across membranes. By coupling solute movement to ion transport, secondary transporters are able to concentrate solutes across a membrane by a factor of 10^6 (ref. 2), and solute flux can occur 10^5 times faster than by simple passive diffusion[3,4].

Secondary transporters are found in all species throughout the kingdoms of life[5]. In humans, they participate in a range of physiological processes, from the uptake of nutrients in the intestine[6] to the transport of Na^+ and Cl^- in the kidney[7] and the removal of neurotransmitters from the synaptic cleft[8]. Secondary transporters are therefore the target of multiple therapeutic agents, including thiazide diuretics, which inhibit a Na^+/Cl^- symporter in the distal convoluted tubule of the kidney[9], and selective serotonin re-uptake inhibitors (antidepressants), which block the activity of the serotonin transporter[10].

At the level of primary structure, analyses of amino-acid sequence suggest that there are more than 100 distinct families of secondary transporter[11], with more than 40 families identified in humans alone[12]. With respect to biological function, these amino-acid sequences encode transporters that act on a range of substrates including elemental cations and anions, aromatic neurotransmitters, nutrients and di- and tripeptides[13,14]. Transport is usually driven by proton or sodium transmembrane gradients[12,13].

In this Review, we focus on transporters that are coupled to sodium ions. We will discuss what recent crystallographic advances relating to sodium-coupled transporters can tell us about the coupling of substrates to ions, the conformational state of the transporter at different stages of the transport cycle, the mechanisms of transport inhibition, and how the substrate and ion pathway is alternately opened and closed — or gated — to maintain a tightly coupled transport mechanism.

Alternating-access mechanism

The mechanism by which secondary transporters couple the chemical potential of an ionic gradient to the translocation of solute has been debated for decades. Peter Mitchell provided early insights into the mechanism of secondary transporters by suggesting that they occupy two alternating structural states: one in which the substrate-binding pocket is accessible to extracellular solution ('open-to-out'), and another in which the binding pocket is accessible to the cytoplasm ('open-to-in')[15]. In this simple model, coupled transport occurs by the synergistic binding of substrate and ion to the open-to-out state followed by isomerization of the transporter to the open-to-in state, allowing the release of both substrate and ion to the cytoplasm[16]. In the late 1950s and 1960s, the basic idea of a two-state alternating-access mechanism was recast in several forms, from the 'gate-type non-carrier' mechanism of Clifford Patlak[17] to the two-state 'allosteric model' of George Vidaver[18] and the 'alternating access' model of Oleg Jardetzky[19].

Molecular mechanisms of secondary transporters based on atomic structures did not emerge until almost 40 years later, largely because these transporters are hydrophobic and dynamic, making them difficult to crystallize. In 2002, the first crystal structure of a secondary transporter was reported[20], the proton-driven multidrug efflux pump AcrB of the resistance nodulation cell division (RND) family from *Escherichia coli*. In 2003, the crystal structures of two major facilitator superfamily (MFS) transporters were solved: the glycerol-3-phosphate/phosphate antiporter GlpT[21] and the proton-coupled lactose symporter LacY[22]. Even though AcrB has a markedly different fold from the MFS transporters, each of these structures revealed an internal two-fold structural pseudo-symmetry that relates the amino-terminal half of the transporter to the carboxy-terminal half by an axis running through the centre of the transporter, approximately perpendicular to the membrane. Furthermore, the outward-facing conformations adopted by GlpT and LacY suggest that the transport mechanism involves a 'rocker switch' motion of the two symmetry-related halves, alternately opening and closing 'gates' to the extracellular and intracellular solutions.

The first atomic-resolution structural insights into the mechanisms of sodium-coupled secondary transporters were reported in 2004 and 2005 with the structures of the aspartate transporter GltPh[23] of the dicarboxylate/amino-acid:cation symporter (DAACS) family, followed by the Na^+/H^+ antiporter NhaA from *E. coli*[24] and the bacterial leucine transporter LeuT[25] of the neurotransmitter:sodium symporter (NSS) family. The structures of GltPh, NhaA and LeuT not only revealed unique membrane protein folds but also reinforced the theme of internal two-fold

[1]Vollum Institute, [2]Department of Biochemistry and Molecular Biology, and [3]Howard Hughes Medical Institute, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Oregon 97239, USA.

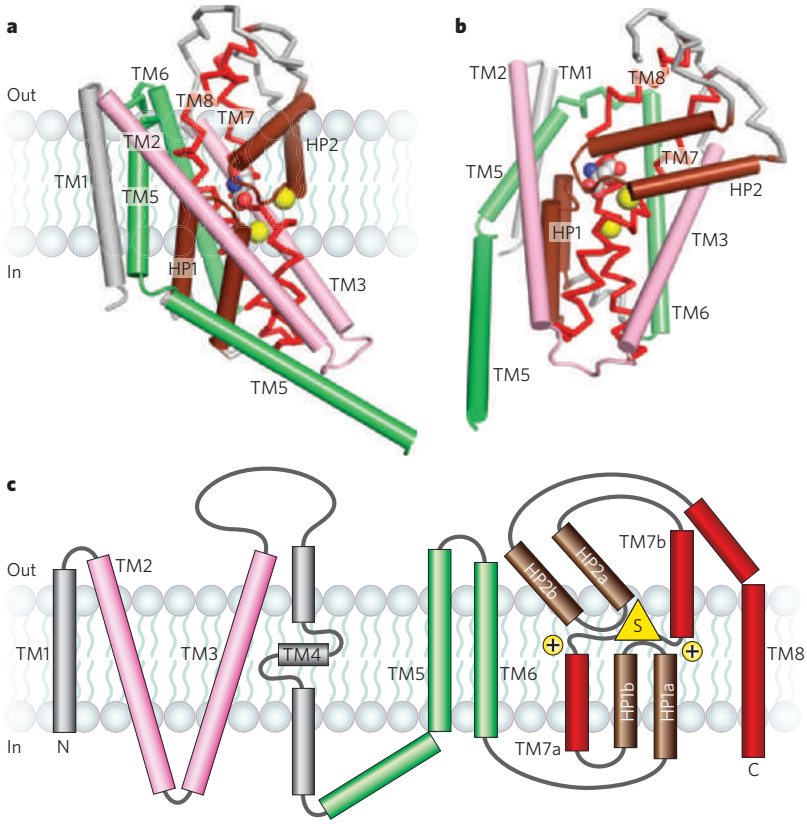


Figure 1 | Architecture of the GltPh fold. **a**, The core transmembrane helices of GltPh are shown, illustrating how the first six transmembrane segments surround the elements of the transporter machinery. The inverted cradle formed by TM2/TM3 and TM5/TM6 is coloured pink and green, respectively. The re-entrant hairpins (HP1 and HP2) are shown in brown and the partly unwound TM7 and the amphipathic TM8 are shown in red. The view is parallel to the membrane and only one subunit of the GltPh trimer is shown. TM4 is omitted for clarity **b**, The same elements as **a** viewed approximately perpendicular to the membrane. The bound substrates (carbon, grey; oxygen, red; nitrogen, blue) and sodium ions (yellow) are shown. **c**, Topology diagram for GltPh with substrate and ions depicted as yellow triangle and circles, respectively.

structural symmetry and discontinuous transmembrane helices[26].

GltPh, which assembles as a homotrimer, displays a pseudo-two-fold symmetrical relationship between crucial elements of the protomer architecture, including two re-entrant hairpin loops (HP1 and HP2), together with the transmembrane helix TM7a and the first half of TM8 (ref. 23). The relevance of the two-fold axis to the transporter mechanism is particularly striking, and it suggests that HP1 and HP2 may undergo alternating, symmetry-related motions that open and close access to the substrate- and ion-binding sites[23] (Fig. 1).

In LeuT, which has a different fold from GltPh, an internal two-fold pseudo-symmetry axis, running parallel to the membrane plane through the centre of the transporter, relates the first five transmembrane helices (TM1–TM5) to the second five helices (TM6–TM10)[25] (Fig. 2). Surprisingly, the fold seen in LeuT was also observed in the subsequently reported structure of the galactose transporter vSGLT of the solute:sodium symporter (SSS) family[27], and in the benzyl-hydantoin transporter Mhp1 of the nucleobase:cation symporter (NCS1) family[28]. Both vSGLT (ref. 27) and Mhp1 (ref. 28) contain the '5+5' inverted structural symmetry motif defined by TM1–TM10 of LeuT, even though these three transporters do not share significant amino-acid sequence identity or have the same number of transmembrane segments. In vSGLT, which has 14 transmembrane helices compared with 12 in LeuT and Mhp1, an N-terminal transmembrane helix precedes the 5+5 helix repeat and three additional helices follow the repeat. That different transporters have the same common helix core but have additional transmembrane segments on the periphery supports the idea that the two-fold-related 5+5 transmembrane repeat defines the fundamental machinery of these transporters.

Not only have the crystal structures of vSGLT, LeuT and Mhp1 effectively 'collapsed' the SSS[29], NSS[30] and NCS1 transporter families into one structural group, they also suggest that other secondary transporters, previously believed to belong to distinct families, may also have LeuT-like folds. The similarity in architecture among LeuT, vSGLT and Mhp1 further implies commonalities in mechanism, ranging from the principles of substrate and ion binding and specificity to conformational changes associated with transport. Atomic models of these functionally disparate yet structurally related transporter families have provided insight into the principles of sodium-coupled transport, and are beginning to clarify an alternating-access mechanism that is distinctly different from that of the MFS family. Conversely, comparison of the structurally disparate LeuT and GltPh transporters suggest commonalities in the concept of 'gates', how they function in the alternating-access mechanism, and in the mechanism of competitive inhibition.

A central pathway inside a scaffold

Close inspection of the LeuT, vSGLT and Mhp1 structures (Fig. 2) shows that the 5+5 transmembrane motif consists of two interior pairs of symmetry-related helices — TM1 and TM6, and TM3 and TM8 — that are nested within an outer ring of helices, TM2, TM4, TM5, TM7, TM9 and TM10 (numbered according to LeuT). Consistent with mutagenesis and functional studies[31–34], these interior pairs largely define the central translocation pathway that contains the binding sites for substrate and ions. The three structures show that the substrate-binding site lies in the centre of the interior pairs, and coincides with the internal two-fold symmetry axis.

Among the outer ring of helices, symmetry-related TM4 and TM5, and TM9 and TM10, form inverted V-shaped pincers that cradle the interior pair of TM3 and TM8, whereas TM2 and TM7, also related by the two-fold axis of symmetry, link TM1 and TM6 with the intracellular and extracellular helix–loop–helix structures, IL1 and EL4. We suggest that the outer ring of helices, which nestles around the interior pairs, provides a framework to stabilize the transporter within the lipid membrane, and that it couples conformational changes on one side of the membrane to movements on the other side.

The central translocation pathway surrounded by a protein scaffold is also observed in the GltPh fold (Fig. 1). In GltPh, the transport machinery of HP1, HP2, TM7 and TM8, forming a C-terminal domain, is enveloped by a ring of six transmembrane helices from the N-terminal domain of the transporter. In this case, the crucial role of the C-terminal domain in defining the transport pathway was suggested by studies showing that functionally important residues were localized to the C terminus, and that the C-terminal domain was more highly con-

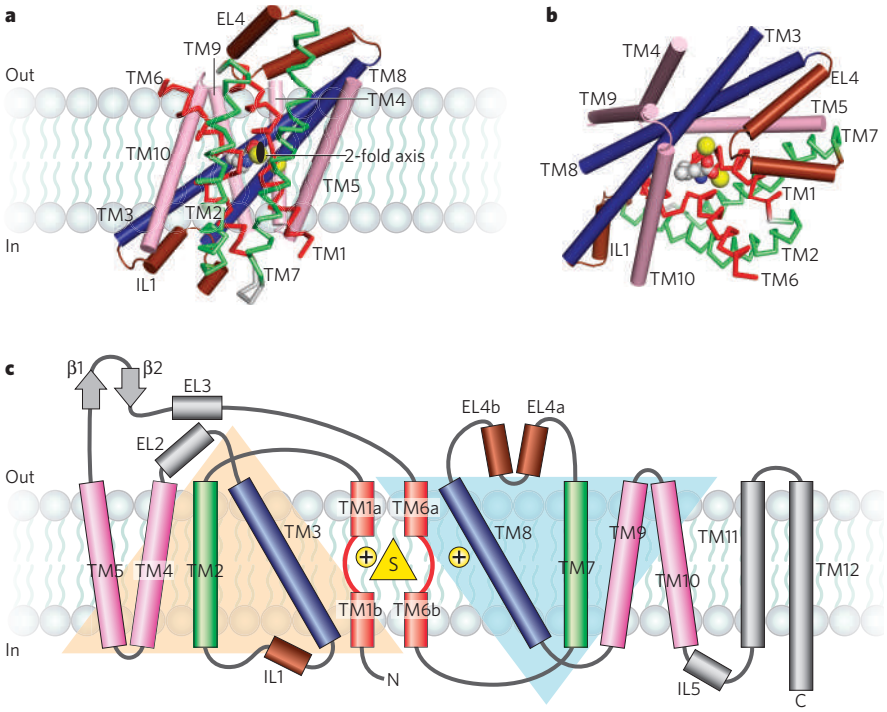


Figure 2 | Architecture of the LeuT fold. **a**, View of the core 5+5 repeat structure for LeuT showing the inverted scaffold of TM4, TM5, TM9 and TM10 holding the long bracing helices of TM3 and TM8 and the jointed, finger-like and partly unwound TM1 and TM6 helices. Bracing TM1 and TM6 are TM2 and TM7 (green). Re-entrant, pseudo-two-fold related loops that either partly (EL4) or fully (IL1) occlude central binding sites are shown in brown. View is parallel to the membrane. **b**, The same elements as in **a** viewed approximately perpendicular to the membrane. The bound substrate (carbon, grey; oxygen, red; nitrogen, blue) and sodium ions (yellow) are shown. **c**, Topology diagram for LeuT with substrate and ions depicted as yellow triangle and yellow circles, respectively. The large beige and blue triangles overlap the five helix repeats related by the pseudo-two-fold axis of symmetry.

served than the N-terminal one[35–37]. For GltPh and its orthologues, the scaffold of TM1–TM6 not only supports elements of the transport pathway, but also mediates essential intersubunit contacts in the trimer.

Substrate- and ion-binding sites

LeuT has a single substrate-binding site at its centre, surrounded by the interior helices, TM1, TM3, TM6 and TM8 (ref. 25). The binding sites for galactose in vSGLT[27] and benzyl-hydantoin in Mhp1[28] are also similarly located (Fig. 3a). Directly adjacent to the primary binding site, TM1 and TM6 in LeuT and Mhp1, or TM2 and TM7 in vSGLT, have interruptions in their helical conformations, a structural feature seen in other membrane proteins that transport ions[24,38,39]. The interruption in the α-helical structures in the proximity of the binding site exposes main-chain hydrogen-bonding partners and orients the helical dipoles to create a polar environment for coordinating substrate and ions within the lipid bilayer[25]. These electrostatic elements, together with side-chain atoms, sculpt the steric, chemical and electrical properties of the binding pocket, conferring on a given transporter selectivity for a substrate based on its size, polarity and charge[40].

Recent ligand-binding experiments and steered-molecular-dynamics simulations of LeuT have suggested that there is an additional secondary binding site between the primary site and the bulk extracellular solution, located near R30 and D404 (refs 41, 42). It has been proposed[41] that the simultaneous occupancy of this secondary site triggers the

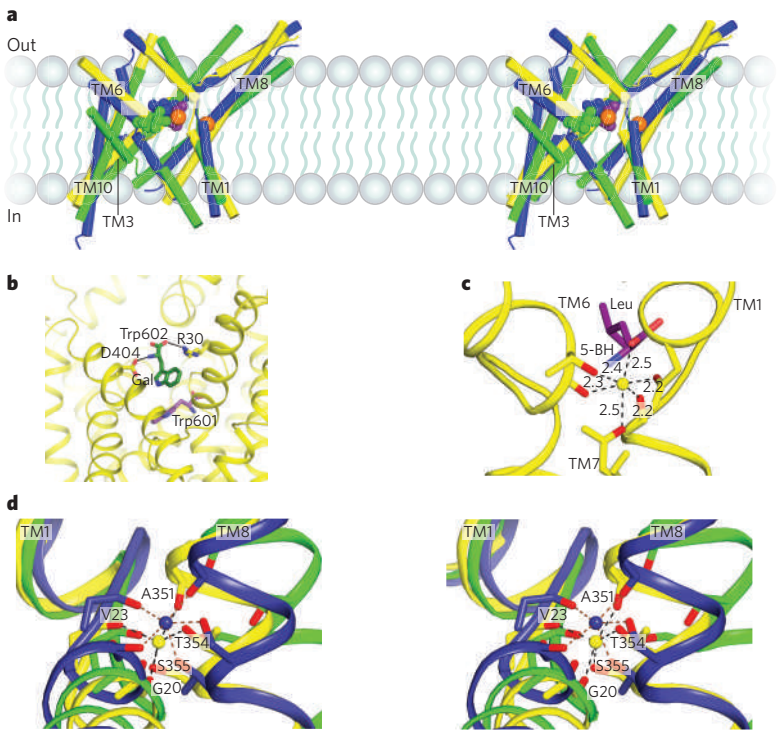


Figure 3 | Conserved substrate- and ion-binding sites in LeuT, vSGLT and Mhp1. **a**, Stereo diagram of the superpositioned occluded structures of LeuT (yellow), Mhp1 (blue) and vSGLT (green) showing the location of their primary substrate-binding sites roughly in the middle of the membrane bilayer and close to the discontinuous regions of TM1 and TM6. Substrate (Gal, galactose (green); 5-BH, 5-benzyl hydantoin (blue); Leu, leucine (purple)) and LeuT Na^+ ions (orange) are shown. For clarity, only TM1, TM3, TM6, TM8 and TM10 are shown. **b**, View of the secondary binding site in LeuT. A second Trp molecule, Trp 602, is bound between R30 and D404 in the open-to-out conformation stabilized by Trp 601 in the primary binding site. **c**, The sodium ion (yellow sphere) at the Na1 site in LeuT is octahedrally coordinated by residues from TM1, TM6 and TM7 as well as bound leucine (purple). Distances in Å are shown. **d**, Stereo representation of superpositioned LeuT (yellow), vSGLT (green) and Mhp1 (blue) structures shows the location of their Na2 binding sites. The sodium ions at the Na2 site of LeuT and Mhp1 are shown as yellow and blue spheres, respectively. Residues contributing side-chain and main-chain oxygens that coordinate the sodium ions are shown as sticks with LeuT residues (A351, T354, S355, G20 and V23) labelled.

Table 1 | Families of sodium-coupled transporters grouped according to structural fold

Structural fold	Family	Transporter	Transported solute	Co-transported ions
LeuT fold	NSS (SLC6)	*LeuT*	Amino acids	$2Na^+$
		TyT1	Tyrosine	$2Na^+$
		TnaT	Tryptophan	$2Na^+$
		CAATCH1	Neutral amino acids	$2Na^+$ or $2K^+$
		CRT	Creatine	$2Na^+$, $1Cl^-$
		GlyT1c	Glycine	$2Na^+$, $1Cl^-$
		GlyT2b	Glycine	$3Na^+$, $1Cl^-$
		NET	Noradrenaline	$1Na^+$
		GAT-1	γ-Aminobutyric acid	$2Na^+$, $1Cl^-$
		SERT	Serotonin	$1Na^+$, $1Cl^-$, $1K^+$
		DAT	Dopamine	$2Na^+$
		$B^{0,+}$	Neutral and cationic amino acids	$2Na^+$, $1Cl^-$
	SSS (SLC5)	*vSGLT*	Glucose/galactose	$1Na^+$
		PutP	Proline	$1Na^+$
		NIS	Iodide	$2Na^+$
		PanF	Pantothenate	$1Na^+$
		SMCT	Monocarboxylate	$3Na^+$
		SMIT2	Myoinositol	$2Na^+$
	NCS1	*Mhp1*	Hydantoin	$1Na^+$
		CodB	Cytosine	$1H^+$
		Nrt1	Nicotinamide riboside	$1H^+$
		Thi10	Thiamine	$1H^+$
		Tpn1	Vitamin B6	$1H^+$
GltPh fold	DAACS (SLC1)	*GltPh*	Aspartate	$1Na^+$
		DctA	C4-dicarboxylate	$2H^+$
		EAAT 1-5	Glutamate/aspartate	$3Na^+$,$1H^+$,$1K^+$
		AscT1	Neutral amino acids	$1Na^+$
		System B^0	Broad-specificity amino acids	$1Na^+$
NhaA fold	NhaA	*NhaA*	Sodium ion	$2H^+$

Families of sodium-coupled transporters (for classification see www.tcdb.org)[14] are grouped into the three known structural folds. Where applicable, the SLC class[12] is indicated in parenthesis. The name of each fold is based on the first transporter structure solved that has that fold. Representative transporters from each family are listed, and ion stoichiometry is indicated in the 'Co-transported ions' column when supported by biochemical data (www.tcdb.org)[14]. Transporters for which crystal structures have been determined are italicized. Some of the transporters are H^+ rather than Na^+ dependent.

intracellular release of substrate and sodium ions from the primary site. X-ray diffraction studies of LeuT, by contrast, do not show binding of either the substrate leucine or the substrate analogue selenomethionine anywhere other than the primary binding site[40]. However, LeuT complexed with tryptophan, which locks the transporter in an open-to-out conformation, does bind a second tryptophan molecule (Trp 602) between R30 and D404 (ref. 40) (Fig. 3b). We suggest that this site is transiently occupied as substrates move from the extracellular vestibule to the primary binding site when the transporter is in the open-to-out conformation.

The high-resolution structure of LeuT also identified the presence of two Na^+ binding sites, which we label as Na1 and Na2 (ref. 25). The sodium ion at the Na1 site is octahedrally coordinated by five protein ligands and the carboxylate of the substrate leucine (Fig. 3c), demonstrating that ion and substrate binding are directly coupled. By contrast, the Na2 site is located roughly 6 Å away from the substrate in LeuT and the sodium ion at this site is bound with a trigonal bipyramidal coordination geometry (Fig. 3d). Intriguingly, by structural comparison, a sodium-ion binding site similar to the Na2 site in LeuT has been identified in both vSGLT and Mhp1, positioned about 10 Å away from the substrates[27,28]. Although the resolutions of the vSGLT and Mhp1 structures are not high enough to unambiguously assign a sodium ion to the site, a sodium ion at this position in vSGLT is supported by biochemical and mutagenesis studies on vSGLT and other SSS family members, including the sodium/iodide symporter[27,43,44]. These observations indicate a role for an ion bound at the Na2 site not only in substrate binding, but also in conformational changes associated with substrate transport. Studies on GAT1 (ref. 45), as well as molecular-dynamic and free-energy simulations of LeuT[46], suggest that the Na2 site is a low-affinity site that can readily give up its ion to the bulk phase, promoting release of the substrate[47].

Sodium-to-substrate stoichiometry varies not only between sodium-coupled transporters, but also between members of the same family, depending on the thermodynamic driving force required for substrate uptake[4]. The requirement for transport varies from one to three sodium ions per substrate in the NSS[2,48–50] and SSS families[44,51–53] (Table 1). Because vSGLT and Mhp1 probably have a sodium-ion binding site similar to the LeuT Na2 site, we suggest that this is a common ion site for divergent transporters and is essential for coupled substrate binding and symport. Although the Na1 site is less conserved among these transporters from different families, a sodium ion bound at this site not only enhances substrate binding for members of the NSS family, but also provides favourable interactions with the co-transported chloride ion[54,55]. Some LeuT orthologues couple substrate transport to three sodium ions, but there is no direct experimental evidence for the location of a third sodium-ion binding site[4].

Conformational states

The crystal structures of LeuT[25,40,56], Mhp1 (ref. 28), vSGLT[27] and GltPh[23,57] provide evidence for the conformations of sodium-coupled secondary transporters as they proceed through the transport cycle. These structures are consistent with a mechanism of transport (Fig. 4a) in which an outward-facing conformation of the transporter (T^{out}) binds substrate and ions and subsequently isomerizes to an inward-facing conformation (T^{in}), via substrate- and ion-bound intermediate states (T^MS^{in} and T^MS^{out}) analogous to enzyme–substrate Michaelis complexes. After

the release of the substrate and ions, the T^{in} state recycles back to the T^{out} state, either in the apo form (without a bound substrate), through a potassium-bound state (as is the case for glutamate[58] and serotonin transporters[59]), or back through a $T^{M}S$ state (with substrate exchange).

One observation from the crystallographic studies is the presence of a stable occluded state for the substrate- and ion-bound ternary complex of each of the four transporters (Fig. 4a). This state is characterized by the bound substrate residing in a closed or partly occluded binding pocket, where dissociation from the pocket would require a conformational change. Despite the common steric occlusion of the substrate, the degree to which the four transporters block solvent accessibility to the binding pocket from the extracellular and cytoplasmic sides varies (Fig. 4a).

In the substrate-bound state of Mhp1 and LeuT, the occluded state has an outward-facing conformation ($T^{M}S^{out}$), with the extracellular pathway being kept open to solvent. In LeuT, the extracellular solvent-exposed region is formed by a large hydrophobic vestibule. At the base of this vestibule are two highly conserved residues, Tyr 108 and Phe 253, which close the top of the binding pocket, creating an occluded substrate-binding site. In Mhp1, the structural elements that occlude the substrate benzyl-hydantoin are different from those of LeuT and involve the N-terminal half of TM10. GltPh also has an outward-facing occluded conformation (Fig. 4a). Aspartate is bound between the tips of the HP1 and HP2 loops, which are closed over the binding site like lids, preventing the dissociation of substrate to either side of the transporter.

In contrast to LeuT and Mhp1, the occluded state of vSGLT adopts an inward-facing conformation ($T^{M}S^{in}$), exposing a cavity to intracellular solution (Fig. 4a), consistent with accessibility studies carried out on PutP[43] and SERT[60–63], orthologues of vSGLT and LeuT, respectively. Akin to leucine binding in LeuT, galactose is bound to vSGLT in a central binding pocket located above the intracellular vestibule, and is occluded from the vestibule by a conserved aromatic residue, Tyr 263. The inward-facing, occluded conformation of galactose-bound vSGLT is structurally different from that observed for substrate-bound LeuT or Mhp1, but there is a simple relationship between the two distinct states: the outward- and inward-facing states are related by the common two-fold axis of internal symmetry that relates the 5+5 transmembrane repeats, suggesting that the symmetrical relationship between LeuT and vSGLT approximates the mechanistic relationship between the $T^{M}S^{in}$ and $T^{M}S^{out}$ states.

The crystal structure of Mhp1 in the unliganded form shows an open-to-out conformation representing the T^{out} state of the transport cycle (Fig. 4a). Comparison of the ligand-bound occluded form of Mhp1 with the apo open-to-out state shows that the N-terminal half of TM10 bends inwards in response to ligand binding to form the occluded state. Further insight into the conformation of the T^{out} state comes from the crystal structures of LeuT and GltPh bound to competitive inhibitors that trap the open-to-out conformations of the transporters[40,57] (Fig. 4b).

Mechanisms of inhibition

The crystal structures of LeuT bound to competitive and non-competitive inhibitors have afforded us a glimpse into the mechanisms of inhibition for the NSS family of transporters. In 2007, crystal structures of LeuT bound to the tricyclic antidepressants (TCAs) clomipramine,

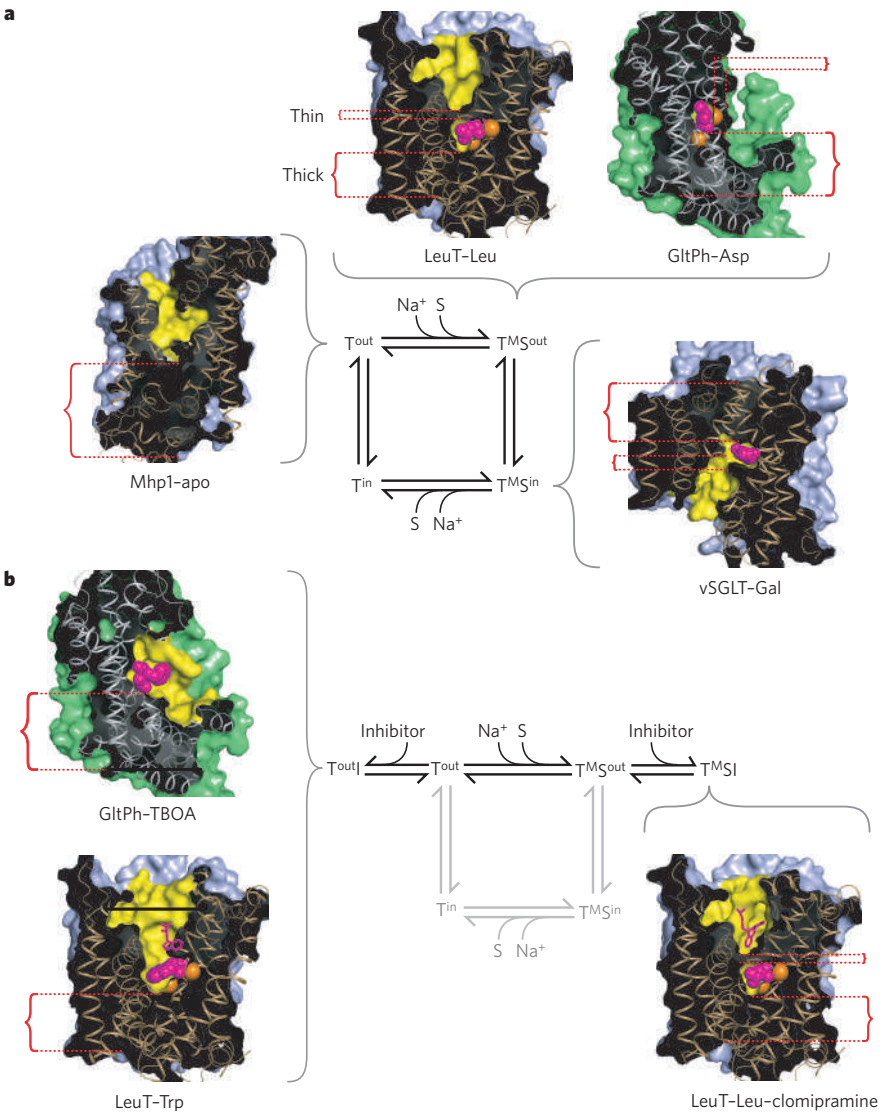


Figure 4 | Crystal structures of transport intermediates. **a**, Transport cycle based on an alternating-access-type mechanism together with insights from crystallographic studies. Clockwise from the T^{out} state: Mhp1–apo (PDB code 2JLN), LeuT–Leu (PDB code 2A65), GltPh–Asp (PDB code 2NWX) and vSGLT–Gal (PDB code 3DH4). **b**, The inhibitory branches of the transport cycle from **a**. On the left, structures of GltPh–TBOA (PDB code 2NWW) and LeuT–Trp (PDB code 3F3A) represent an open-to-out competitive inhibitor-bound state. On the right, the structure of LeuT–Leu–clomipramine (PDB code 2Q6H) represents a non-competitive inhibitor-bound occluded state. Cross-sectional illustrations of the crystal structures of each transporter are shown associated with the states of the cycle they represent. The positions of the 'thin' gates and 'thick' gates are highlighted by red dashed lines. The solvent-accessible surface area, calculated with a probe radius of 1.4 Å, is shown in light blue for the LeuT-fold structures and green for GltPh-fold structures. Yellow regions highlight the surfaces of the binding site and cavities that penetrate the structures. Bound ligands, shown as van der Waals spheres, are coloured magenta, with sodium ions in orange. The view of each transporter is approximately parallel to the membrane plane, with the extracellular side at the top of each figure. PDB, Protein Data Bank; TBOA, aspartate analogue.

desipramine and imipramine were reported[56,64]. These molecules have therapeutic value as competitive inhibitors of the human serotonin transporter[65] and block re-uptake of serotonin from synapses, thereby prolonging activation of the serotonin receptor. For LeuT, however, the mechanism of inhibition by TCAs is non-competitive[56]. The structures of the LeuT–TCA complexes reveal that the TCA molecule binds in the outward-facing vestibule, a partly hydrophobic cavity that binds other non-polar molecules, including *n*-octyl-β-D-glucopyranoside[40]. The TCA is situated directly above the R30–D404 salt bridge, where the guanidinium head group of the arginine has flipped to form a salt bridge with the aspartate, stabilizing the occluded state of LeuT ($T^{M}SI$; Fig. 4b), and preventing the further conformational changes needed for progress around the transport cycle. The identification of this inhibitory allosteric site is consistent with a general mechanism of non-competitive inhibition in which the substrate-binding site and inhibitor site do not overlap, trapping the transporter in an inactive state. The non-competitive mechanism for the inhibition of LeuT by TCAs is different from the competitive mechanism for the inhibition of SERT by TCAs[65]. Nevertheless, the structural principles revealed by the LeuT–TCA complexes define a mechanism of allosteric inhibition in NSS family transporters and, by extension, in other transporters with the LeuT fold.

The structural basis for competitive inhibition was recently revealed by the crystal structure of LeuT bound to tryptophan[40]. Tryptophan acts as a strut, where the bulky indole ring is wedged into the binding pocket, displacing the α-amino and α-carboxylate moieties outwards by about 2 Å relative to their positions in the leucine-bound occluded state. There is insufficient space to fully accommodate the indole ring in the substrate-binding pocket, so the transporter is effectively propped open by interactions of the inhibitor's α-substituents with TM1b and TM6a, and of the indole ring with TM3, TM8 and TM10. In this way, the transporter is locked open ($T^{out}I$; Fig. 4b) and blocked from progressing to the occluded $T^{M}S^{out}$ state of the transport cycle.

The crystal structure of GltPh bound to the competitive inhibitor TBOA, a bulky aspartate analogue, reveals a similar principle of competitive inhibition for transporters with the GltPh fold[57] (Fig. 4b). The aspartate group of TBOA binds in a similar position as the substrate L-aspartate, lodged between TM7, TM8 and HP1. However, the large benzyl moiety of TBOA sticks out towards HP2, propping HP2 in an open conformation (Fig. 4b), disrupting sodium site 2 and precluding the formation of the occluded state.

The observation that bulky substrate analogues act as competitive inhibitors of transport to stabilize an opening of the extracellular side of the transporter is supported by substituted cysteine accessibility method (SCAM) assays for the eukaryotic NSS homologue GAT1, as well as the human glucose transporter hSGLT, a member of the SSS family[66]. The consistency of these results with the LeuT–Trp crystal structure suggests that the mechanism of inhibition is likely to be similar for other sodium-coupled transporters that share the LeuT fold, and that a comparable principle seems to be found in other families of structurally disparate transporters, such as those adopting the GltPh fold.

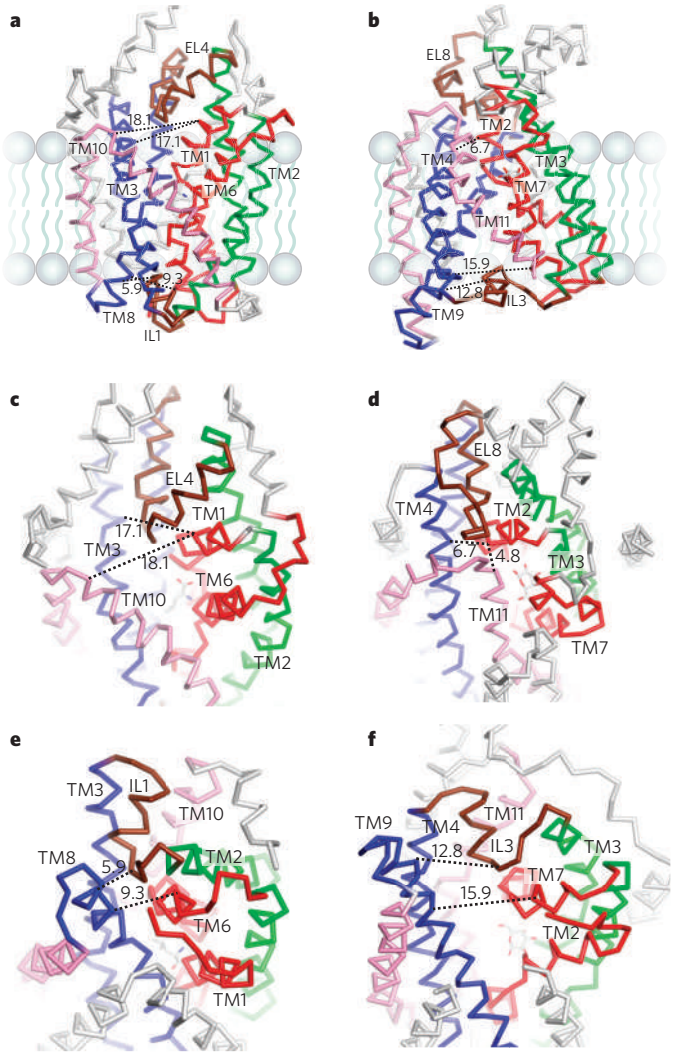


Figure 5 | Comparative views of substrate-bound LeuT in the $T^{M}S^{out}$ state (left) and vSGLT in the $T^{M}S^{in}$ state (right). **a**, **b**, Views parallel to the membrane of LeuT (**a**, PDB code 2A65) and vSGLT (**b**, PDB code 3DH4). **c**, **d**, Top-down view of the extracellular pathway for LeuT (**c**) and vSGLT (**d**). **e**, **f**, Bottom-up view of the intracellular pathway for LeuT (**e**) and vSGLT (**f**). Equivalent structural elements are coloured the same in both LeuT and vSGLT. To help gauge the re-organization of the extra- and intracellular elements, black dashed lines indicate distances between structural elements, measured from structurally similar residues in the two transporters. Considering the internal two-fold symmetry, note the similar organization of the open LeuT extracellular pathway (**c**) to the open vSGLT intracellular pathway (**f**), and of the closed vSGLT extracellular pathway (**d**) to the closed LeuT intracellular pathway (**e**). PDB, Protein Data Bank.

Permeation pathways and gating mechanisms

In transporters with the GltPh fold (GltPh) and the LeuT fold (LeuT, vSGLT and Mhp1), the primary substrate- and ion-binding sites are flanked by two gates, one controlling access to the outside of the cell, the other controlling access to the inside. Only one of these gates can open at a time, allowing substrates and ions to reach the primary binding sites without opening up a continuous transmembrane pore. Understanding how secondary transporters work is fundamentally a question of how the gates work: what principles govern the coordinated, alternate opening and closing of the extracellular and intracellular gates when substrates bind on the outside and unbind on the inside. To answer this question, we must consider the conformational changes that occur during transport and the likely pathways that substrates and ions take when they bind to, and unbind from, their primary sites.

In the small group of sodium-coupled transporter structures, for a given transporter trapped in a specific state, the gates that control access to and from the primary binding site are often asymmetric, with the extracellular gate being less substantial or 'thinner' than the cytoplasmic gate, or vice versa (Fig. 4a). This is observed for transporters with the GltPh fold as well as those with the LeuT fold. For example, in the outward-facing occluded leucine-bound LeuT complex, only a few residues directly block access from the primary binding site (Tyr 108 and Phe 253), forming a 'thin' gate at the base of a solvent-filled cavity to the outside. By contrast, the cytoplasmic 'thick' gate is made up of about 20 Å of packed protein, including TM1a, TM3, TM6b, TM8 and TM10, along with the N terminus and IL1 (Fig. 5a, c, e). Similarly, in the substrate-bound state of GltPh, the extracellular gate is made up of just a few residues at the tip of HP2, whereas the cytoplasmic gate is composed of a roughly 15-Å slab of helices and side chains (HP1, TM7a and TM8).

The substrate- and ion-bound inward-facing occluded state of

vSGLT presents a converse situation in which a thick extracellular gate is formed by TM1b, TM3, TM6a, TM10 and EL4 (numbering as LeuT) and a thin cytoplasmic gate is composed of Tyr 262, Tyr 263 and Trp 264 (Fig. 5b, d, f). Thus, the thin gates are typically defined by the side-chain atoms of a few residues, whereas the thick gates are formed by transmembrane helices packed close together, in combination with extracellular and intracellular loops such as N termini, IL1 or EL4. A prominent observation from the structures of outward-facing LeuT and Mhp1, and inward-facing vSGLT, is that the thick gate of each transporter is related to the extracellular or intracellular pathway by the internal two-fold symmetry. In any one state, the structural components of the solvent-filled pathway are reciprocal (with respect to the axis of the two-fold symmetry) to those of the thick gates.

Insight into how the thin gate opens and closes is provided by the structures of LeuT, Mhp1 and GltPh captured in both open-to-out (T^{out}) and outward-facing occluded ($T^{M}S^{out}$) conformations. These structures demonstrate that substrate and ion binding results in relatively small conformational changes. In GltPh, for example, aspartate binding allows HP2 to close over the binding site, whereas TBOA binding holds it open, suggesting that the simple 'flipping' movement of HP2 primarily describes the motion of the thin gate that occludes the binding site during transport. With LeuT, in comparing the open-to-out conformation of the Trp complex with the occluded Leu-bound state, the most substantial change is the rotation of a subdomain of the transporter composed of TM1b, TM2a, TM6b and EL4, which, together with rotations of several side chains, collectively moves inwards to close the thin gate, occluding the substrate-binding site from extracellular solution. For Mhp1, the binding of substrate involves the inward bending of the N-terminal half of TM10. Thus the thin gate opens and closes around the substrate-binding pocket upon substrate binding or unbinding, defining the transitions from T^{out} to $T^{M}S^{out}$ or from $T^{M}S^{in}$ to T^{in} states .

In contrast to the local changes associated with substrate binding, the isomerization between the outward-facing ($T^{M}S^{out}$) and inward-facing ($T^{M}S^{in}$) states involves larger-scale conformational changes spread throughout the transporter. These conformational changes can be conceptualized by applying the two-fold axis of internal pseudo-symmetry to the key transmembrane helices TM1 and TM6, which deviate from this symmetry[25,60]. Rotation of TM1 and TM6 of LeuT about the internal symmetry axis generates an inward-facing model in which the extracellular gate is closed and the intracellular gate is open[60]. Based on this model, it was suggested that the bundle of TM1, TM2, TM6 and TM7 moves as a rigid body in a rocker-switch-like mechanism to alternately open and close the extracellular and intracellular gates[60]. However, this simplification is not consistent with the structural comparisons of the LeuT–Leu and LeuT–Trp complexes or of the Mhp1–apo and substrate-bound states, which indicate that, for example, TM1 is not a rigid body and that there is some degree of independent movement within the helix bundle. Further experimental and computational studies are required to understand the movements that describe this conformational change.

Nevertheless, comparison of the LeuT–Leu (outward-facing, $T^{M}S^{out}$) and vSGLT–Gal (inward-facing, $T^{M}S^{in}$) structures suggests that the differences between these states can be described by a reorientation of TM1 and TM6 (TM2 and TM7 in vSGLT), together with movement and bending of TM2 and TM7 (TM3 and TM8 in vSGLT). In the outward-facing state, near the extracellular opening in LeuT, TM1 is about 17 Å and 18 Å away from TM3 and TM10, respectively (Fig. 5a, c) compared with just 7 Å and 5 Å for the equivalent elements in vSGLT (Fig. 5b, d). Similarly, the intracellular cavity is open in vSGLT about 16 Å, measured between TM9 and TM7 (Fig. 5b, f), whereas the same elements in LeuT (TM6 and TM8), with the thick intracellular gate closed, are about 9 Å apart (Fig. 5a, e). The similar magnitude to which the extracellular cavity of LeuT collapses to form the thick gate seen in vSGLT, and to which the thick intracellular gate of LeuT opens to form the cavity in vSGLT, supports the idea that the relationship between the cavities and the thick gates is reciprocally related by the two-fold internal symmetry of the transporter, and that structures of LeuT and

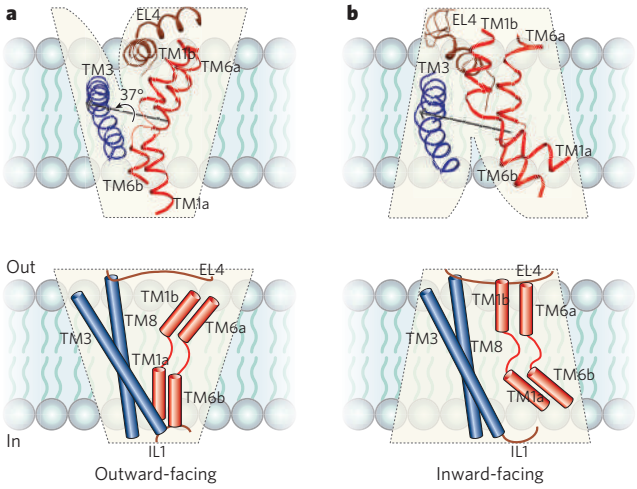


Figure 6 | Transition between outward-facing and inward-facing states in LeuT-fold transporters. Transmembrane segments TM1, TM3, TM6 and TM8 line the central translocation pathway, with EL4 and IL1 acting as lids that seal the extracellular and intracellular gates, respectively, in their closed states. **a**, The outward-facing arrangement of central helices in substrate-bound LeuT. **b**, The inward-facing arrangement of central helices in substrate-bound vSGLT. TM8 and IL1 are omitted from the top section for clarity. TM1 and TM6 rotate approximately 37° relative to TM3 and TM8 in transitioning from the outward-facing state adopted by LeuT (**a**) to the inward-facing state adopted by vSGLT (**b**). The rotation axis, shown in black, and the angle of rotation were calculated using DynDom[67]. Cartoon representations of outward-facing and inward-facing states are shown below the corresponding ribbon diagrams. The cartoon representations are adapted, with permission, from ref. 25.

vSGLT largely represent distinct occluded-state ternary intermediates that interconvert during transport. The reorientation of TM1, TM2, TM6 and TM7 (LeuT numbering) between an occluded LeuT-like conformation ($T^{M}S^{out}$) and an occluded vSGLT-like conformation ($T^{M}S^{in}$) is therefore likely to approximate the conformational transition that reorients the thin gates of a transporter to the opposite side of the membrane. Additionally, the flexing of TM3 and TM8 may also contribute to the opening and closing of the gates, with these transmembrane helices bending at conserved glycine residues near their midsections, and with IL1, EL4 and the N terminus functioning as flexible flaps, helping to seal the gates in the closed state.

Taken together, the crystal structures of the LeuT fold, namely Mhp1, LeuT and vSGLT, identify two major classes of transition that occur during transport. First, substrate binding and unbinding closes and opens, respectively, the thin gates to occlude or expose the substrate in the primary binding site. Second, the opening and closing of the thick gates switches the transporter between outward-facing and inward-facing states. The opening and closing of the thin gates stem from local conformational changes, some of which involve helix rotations centred on axes that pass through the regions of helical discontinuity. By contrast, the thick-gate transition reorients the occluded substrate–transporter complex by rotating the entire membrane-spanning bundle of helices about a central axis approximately perpendicular to the axis of internal two-fold symmetry (Fig. 6a, b).

How do we know that comparing different transporters in different conformations reliably predicts common mechanistic principles? We don't, but the fact that transporters with the LeuT fold share several common elements of structure and symmetry suggests that basic mechanistic principles are also likely to be shared. However, specific details, related to substrate and transporter interactions and regulation, are likely to differ.

What prevents both gates from opening simultaneously? We suggest that the discontinuous helical regions of TM1 and TM6 provide a hinge around which a small degree of conformational change can occur.

This can be seen in the movements that accompany the binding of the competitive inhibitor tryptophan to LeuT, in which the thin extracellular gate opens by outward movements of TM1b and TM6a. However, TM1 and TM6 are adjacent to TM2 and TM7, and together they form a four-helix bundle. Larger-scale movement of TM1b and TM6a is therefore constrained by TM2 and TM7, perhaps because the latter are continuous α-helices that lack the non-helical, hinge-like regions present in TM1 and TM6. So substantial outward (opening) movements of TM1b and TM6a, or of TM1a and TM6b, are limited by TM2 and TM7. As a result, both gates may be closed at the same time, but they are prevented from being simultaneously open by the conformational rigidity enforced by TM2 and TM7.

Future prospects

Recent crystallographic studies of sodium-coupled secondary transporters have greatly advanced our understanding of the structural principles that underlie transporter function. The consistency of these models with the earlier functional studies has allowed us to associate specific conformations with different mechanistic states of the transport cycle. However, this mechanistic description is derived from a patchwork of different transporters fortuitously crystallized in different states. An accurate description of the precise conformational changes that a given transporter undergoes during transport must await further structural, biophysical and computational studies of individual secondary transporters. Additionally, various questions regarding the fundamental nature of the transport cycle remain outstanding. Although it is fairly easy to see how substrate binding leads to the closure of a thin gate, it is harder to deduce the chemical and structural principles that drive the isomerization of the transporter from an outward-facing to an inward-facing state (in other words, open a thick gate). Unlike mechanical models of gating in primary transporters and ion channels, in secondary transporters there is no apparent source of mechanical force to open the thick gate. What structural changes occur in response to the binding of ions? What is the sequence of events that leads to the release of substrate on the cytoplasmic side? How do ions, such as potassium, promote the isomerization of glutamate and serotonin transporters from inward-facing to outward-facing states? Finally, in order to fully understand and appreciate the biological and pharmacological properties unique to human secondary transporters, we will need to solve the crystal structures of eukaryotic homologues. There is clearly much work still to do. ■

1. Quick, M. W. (ed.) *Transmembrane Transporters* (Wiley-Liss, 2002).
2. Roux, M. J. & Supplisson, S. Neuronal and glial glycine transporters have different stoichiometries. *Neuron* **25,** 373–383 (2000).
3. Chakrabarti, A. C. & Deamer, D. W. Permeability of lipid bilayers to amino acids and phosphate. *Biochim. Biophys. Acta* **1111,** 171–177 (1992).
4. Supplisson, S. & Roux, M. J. Why glycine transporters have different stoichiometries. *FEBS Lett.* **529,** 93–101 (2002).
5. Sobczak, I. & Lolkema, J. S. Structural and mechanistic diversity of secondary transporters. *Curr. Opin. Microbiol.* **8,** 161–167 (2005).
6. Wright, E. M. & Turk, E. The sodium/glucose cotransport family SLC5. *Pflügers Arch.* **447,** 510–518 (2004).
7. Hebert, S. C., Mount, D. B. & Gamba, G. Molecular physiology of cation-coupled Cl^- cotransport: the SLC12 family. *Pflügers Arch.* **447,** 580–593 (2004).
8. Sonders, M. S., Quick, M. & Javitch, J. A. How did the neurotransmitter cross the bilayer? A closer view. *Curr. Opin. Neurobiol.* **15,** 296–304 (2005).
9. Gamba, G. Molecular physiology and pathophysiology of electroneutral cation-chloride cotransporters. *Physiol. Rev.* **85,** 423–493 (2005).
10. Murphy, D. L., Lerner, A., Rudnick, G. & Lesch, K. P. Serotonin transporter: gene, genetic disorders, and pharmacogenetics. *Mol. Interv.* **4,** 109–123 (2004).
11. Busch, W. & Saier, M. H. Jr The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol.* **37,** 287-337 (2002).
12. Hediger, M. A. *et al.* The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflügers Arch.* **447,** 465–468 (2004).
13. Saier, M. H. Jr A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.* **64,** 354–411 (2000).
14. Saier, M. H. Jr, Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34,** D181–D186 (2006).
References 11, 13 and 14 provide a comprehensive classification system for membrane transport proteins that includes ion channels as well as primary and secondary transporters.
15. Mitchell, P. A general theory of membrane transport from studies of bacteria. *Nature* **180,** 134–136 (1957).
16. Mitchell, P. in *Advances in Enzymology and Related Areas of Molecular Biology* (ed. Nord, F. F.) 33–87 (Interscience, 1967).
17. Patlak, C. S. Contributions to the theory of active transport: II. The gate type non-carrier mechanism and generalizations concerning tracer flow, efficiency, and measurement of energy expenditure. *Bull. Math. Biophys.* **19,** 209–235 (1957).
18. Vidaver, G. A. Inhibition of parallel flux and augmentation of counter flux shown by transport models not involving a mobile carrier. *J. Theor. Biol.* **10,** 301–306 (1966).
19. Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **211,** 969–970 (1966).
20. Murakami, S., Nakashima, R., Yamashita, E. & Yamaguchi, A. Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature* **419,** 587–593 (2002).
21. Huang, Y., Lemieux, M. J., Song, J., Auer, M. & Wang, D. N. Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science* **301,** 616–620 (2003).
22. Abramson, J. *et al.* Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301,** 610–615 (2003).
23. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
24. Hunte, C. *et al.* Structure of a Na^+/H^+ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435,** 1197–1202 (2005).
25. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature* **437,** 215–223 (2005).
26. Screpanti, E. & Hunte, C. Discontinuous membrane helices in transport proteins and their correlation with function. *J. Struct. Biol.* **159,** 261–267 (2007).
27. Faham, S. *et al.* The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/sugar symport. *Science* **321,** 810–814 (2008).
28. Weyand, S. *et al.* Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science* **322,** 709–713 (2008).
29. Wright, E. M., Loo, D. D. F., Hirayama, B. A. & Turk, E. Surprising versatility of Na^+-glucose cotransporters: SLC5. *Physiology* **19,** 370–376 (2004).
30. Chen, N. H., Reith, M. E. & Quick, M. W. Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6. *Pflügers Arch.* **447,** 519–531 (2004).
31. Barker, E. L., Moore, K. R., Rakhshan, F. & Blakely, R. D. Transmembrane domain I contributes to the permeation pathway for serotonin and ions in the serotonin transporter. *J. Neurosci.* **19,** 4705–4717 (1999).
32. Rudnick, G. Serotonin transporters — structure and function. *J. Membr. Biol.* **213,** 101–110 (2006).
33. Kanner, B. I. & Zomot, E. Sodium-coupled neurotransmitter transporters. *Chem. Rev.* **108,** 1654–1668 (2008).
34. Ben-Yona, A. & Kanner, B. I. Transmembrane domain 8 of the γ-aminobutyric acid transporter GAT-1 lines a cytoplasmic accessibility pathway into its binding pocket. *J. Biol. Chem.* **284,** 9727–9732 (2009).
35. Slotboom, D. J., Lolkema, J. S. & Konings, W. N. Membrane topology of the C-terminal half of the neuronal, glial, and bacterial glutamate transporter family. *J. Biol. Chem.* **271,** 31317–31321 (1996).
36. Grunewald, M., Bendahan, A. & Kanner, B. I. Biotinylation of single cysteine mutants of the glutamate transporter GLT-1 from rat brain reveals its unusual topology. *Neuron* **21,** 623–632 (1998).
37. Seal, R. P. & Amara, S. G. A reentrant loop domain in the glutamate carrier EAAT1 participates in substrate binding and translocation. *Neuron* **21,** 1487–1498 (1998).
38. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405,** 647–655 (2000).
39. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* **415,** 287–294 (2002).
40. Singh, S. K., Piscitelli, C. L., Yamashita, A. & Gouaux, E. A competitive inhibitor traps LeuT in an open-to-out conformation. *Science* **322,** 1655–1661 (2008).
This paper provides a crystallographic and functional analysis of LeuT with several different ligands that correlates the occluded state to transportable ligands and identifies a tryptophan as a competitive inhibitor that traps an open-to-out conformation.
41. Shi, L., Quick, M., Zhao, Y., Weinstein, H. & Javitch, J. A. The mechanism of a neurotransmitter:sodium symporter-inward release of Na^+ and substrate is triggered by a substrate in a second binding site. *Mol. Cell* **30,** 667–677 (2008).
42. Celik, L., Schiott, B. & Tajkhorshid, E. Substrate binding and formation of an occluded state in the leucine transporter. *Biophys. J.* **94,** 1600–1612 (2008).
43. Hilger, D., Bohm, M., Hackmann, A. & Jung, H. Role of Ser-340 and Thr-341 in transmembrane domain IX of the Na^+/proline transporter PutP of *Escherichia coli* in ligand binding and transport. *J. Biol. Chem.* **283,** 4921–4929 (2008).
44. De La Vieja, A., Reed, M. D., Ginter, C. S. & Carrasco, N. Amino acid residues in transmembrane segment IX of the Na^+/I^- symporter play a role in its Na^+ dependence and are critical for transport activity. *J. Biol. Chem.* **282,** 25290–25298 (2007).
The mutation of conserved serine and threonine residues in TM9 of the NIS transporter (SSS family) shows that they are involved in Na^+ binding and translocation.
45. Zhou, Y., Zomot, E. & Kanner, B. I. Identification of a lithium interaction site in the GABA transporter GAT-1. *J. Biol. Chem.* **281,** 22092–22099 (2006).
46. Noskov, S. Y. & Roux, B. Control of ion selectivity in LeuT: two Na^+ binding sites with two different mechanisms. *J. Mol. Biol.* **377,** 804–818 (2008).
47. Caplan, D. A., Subbotina, J. O. & Noskov, S. Y. Molecular mechanism of ion-ion and ion-substrate coupling in the Na^+-dependent leucine transporter LeuT. *Biophys. J.* **95,** 4613–4621 (2008).
48. Talvenheimo, J., Fishkes, H., Nelson, P. J. & Rudnick, G. The serotonin transporter-imipramine "receptor". *J. Biol. Chem.* **258,** 6115–6119 (1983).
49. Gu, H. H., Wall, S. & Rudnick, G. Ion coupling stoichiometry for the norepinephrine transporter in membrane vesicles from stably transfected cells. *J. Biol. Chem.* **271,** 6911–6916 (1996).
50. Keynan, S. & Kanner, B. I. γ-Aminobutyric acid transport in reconstituted preparations from rat brain: coupled sodium and chloride fluxes. *Biochemistry* **27,** 12–17 (1988).

51. Diez-Sampedro, A., Eskandari, S., Wright, E. M. & Hirayama, B. A. Na^+-to-sugar stoichiometry of SGLT3. *Am. J. Physiol. Renal Physiol.* **280,** F278–F282 (2001).
52. Mackenzie, B., Loo, D. D. F. & Wright, E. M. Relationships between Na^+/glucose cotransporter (SGLT1) currents and fluxes. *J. Membr. Biol.* **162,** 101–106 (1998).
53. Eskandari, S. *et al.* Thyroid Na^+/I- symporter. Mechanism, stoichiometry, and specificity. *J. Biol. Chem.* **272,** 27230–27238 (1997).
54. Forrest, L. R., Tavoulari, S., Zhang, Y. W., Rudnick, G. & Honig, G. Identification of a chloride ion binding site in Na^+/Cl^- -dependent transporters. *Proc. Natl Acad. Sci. USA* **104,** 12761–12766 (2007).
55. Zomot, E. *et al.* Mechanism of chloride interaction with neurotransmitter:sodium symporters. *Nature* **449,** 726–730 (2007).
56. Singh, S., Yamashita, A. & Gouaux, E. Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature* **448,** 952–956 (2007).
This paper presents a structural concept for the non-competitive inhibition of NSS transporters by tricyclic antidepressants.
57. Boudker, O., Ryan, R., Yernool, D., Shimamoto, K. & Gouaux, E. Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445,** 387–393 (2007).
58. Kavanaugh, M. P., Bendahan, A., Zerangue, N., Zhang, Y. & Kanner, B. I. Mutation of an amino acid residue influencing potassium coupling in the glutamate transporter GLT-1 induces obligate exchange. *J. Biol. Chem.* **272,** 1703–1708 (1997).
59. Keyes, S. R. & Rudnick, G. Coupling of transmembrane proton gradients to platelet serotonin transport. *J. Biol. Chem.* **257,** 1172–1176 (1982).
60. Forrest, L. R. *et al.* Mechanism for alternating access in neurotransmitter transporters. *Proc. Natl Acad. Sci. USA* **105,** 10338–10343 (2008).
This paper proposes a model for an inward-facing conformation of LeuT based on the structure of outward-facing LeuT–Leu and the inverted structural pseudo-symmetry.
61. Jacobs, M. T., Zhang, Y. W., Campbell, S. D. & Rudnick, G. Ibogaine, a noncompetitive inhibitor of serotonin transport, acts by stabilizing the cytoplasm-facing state of the transporter. *J. Biol. Chem.* **282,** 29441–29447 (2007).
62. Zhang, Y.-W. & Rudnick, G. The cytoplasmic substrate permeation pathway of serotonin transporter. *J. Biol. Chem.* **281,** 36213–36220 (2006).
63. Zhang, Y. W. & Rudnick, G. Cysteine-scanning mutagenesis of serotonin transporter intracellular loop 2 suggests an α-helical conformation. *J. Biol. Chem.* **280,** 30807–30813 (2005).
64. Zhou, Z. *et al.* LeuT-desipramine structure reveals how antidepressants block neurotransmitter uptake. *Science* **317,** 1390–1393 (2007).
65. Apparsundaram, S., Stockdale, D. J., Henningsen, R. A., Milla, M. E. & Martin, R. S. Antidepressants targeting the serotonin reuptake transporter act via a competitive mechanism. *J. Pharmacol. Exp. Ther.* **327,** 982–990 (2008).
66. Hirayama, B. A., Diez-Sampedro, A. & Wright, E. M. Common mechanisms of inhibition for the Na^+/glucose (hSGLT1) and Na^+/Cl^-/GABA (hGAT1) cotransporters. *Br. J. Pharmacol.* **134,** 484–495 (2001).
This paper demonstrates a common principle for inhibition between two functionally unrelated families that hints at the underlying structural conservation.
67. Hayward, S. & Berendsen, H. J. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* **30,** 144–154 (1998).

Acknowledgements This work was supported by the US National Institutes of Health. E.G. is an investigator with the Howard Hughes Medical Institute. We thank R. Hibbs, K. Hollenstein, H. Owen, S. Singh and A. Sobolevsky for helpful comments and L. Vaskalis for assistance with the figures.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to E.G. (gouauxe@ohsu.edu).

NATURE|Vol 459|21 May 2009|doi:10.1038/nature08144

The structure and function of G-protein-coupled receptors

Daniel M. Rosenbaum[1], Søren G. F. Rasmussen[1] & Brian K. Kobilka[1]

G-protein-coupled receptors (GPCRs) mediate most of our physiological responses to hormones, neurotransmitters and environmental stimulants, and so have great potential as therapeutic targets for a broad spectrum of diseases. They are also fascinating molecules from the perspective of membrane-protein structure and biology. Great progress has been made over the past three decades in understanding diverse GPCRs, from pharmacology to functional characterization *in vivo*. Recent high-resolution structural studies have provided insights into the molecular mechanisms of GPCR activation and constitutive activity.

The past two years have seen remarkable advances in the structural biology of G-protein-coupled receptors (GPCRs). Highlights have included solving the first crystal structures of ligand-activated GPCRs—the human β_2 adrenergic receptor (β_2AR), the avian β_1AR and the human A_{2A} adenosine receptor—as well as the structures of opsin and an active form of rhodopsin. These successes followed decades of effort by many laboratories across the world, and are of great interest from the perspectives of membrane-protein biophysics, cell biology, physiology and drug discovery.

GPCRs are the largest family of membrane proteins and mediate most cellular responses to hormones and neurotransmitters, as well as being responsible for vision, olfaction and taste. At the most basic level, all GPCRs are characterized by the presence of seven membrane-spanning α-helical segments separated by alternating intracellular and extracellular loop regions. GPCRs in vertebrates are commonly divided into five families on the basis of their sequence and structural similarity[1]: rhodopsin (family A), secretin (family B), glutamate (family C), adhesion and Frizzled/Taste2. The rhodopsin family is by far the largest and most diverse of these families, and members are characterized by conserved sequence motifs that imply shared structural features and activation mechanisms. Despite these similarities, individual GPCRs have unique combinations of signal-transduction activities involving multiple G-protein subtypes, as well as G-protein-independent signalling pathways and complex regulatory processes. Despite intensive academic and industrial research efforts over the past three decades, little is known about the structural basis of GPCR function. The crystal structures obtained in the past two years provide the first opportunity to understand how protein structure dictates the unique functional properties of these complex signalling molecules.

In this Review, we discuss the similarities and differences among the four known three-dimensional structures of GPCRs in their inactive states. The extracellular surfaces of these structures reveal the molecular underpinnings of antagonist and inverse-agonist ligand recognition. Differences in interactions involving highly conserved residues at the cytoplasmic surface help to explain the varying levels of agonist-independent basal G-protein coupling activity, or 'constitutive activity', among the receptors. We then discuss the recently obtained structures of opsin, which reveal in molecular detail several of the key conformational changes associated with GPCR activation. Finally, we address some of the remaining challenges in the structural biology of GPCRs that must be addressed to fully understand the molecular basis for the physiological function of these proteins.

Multifaceted functionality

Much of vertebrate physiology is based on GPCR signal transduction. As the receptors for hormones, neurotransmitters, ions, photons and other stimuli, GPCRs are among the essential nodes of communication between the internal and external environments of cells. The classical role of GPCRs is to couple the binding of agonists to the activation of specific heterotrimeric G proteins, leading to the modulation of downstream effector proteins. Taking the human β_2AR as an example, the binding of adrenaline and noradrenaline to cells in the target tissues of sympathetic neurotransmission leads to the activation of the stimulatory subunit of the heterotrimeric G protein (Gαs), the stimulation of adenylyl cyclase, the accumulation of cyclic AMP (cAMP), the activation of cAMP-dependent protein kinase A (PKA) and the phosphorylation of proteins involved in muscle-cell contraction[2] (Fig. 1). However, a wealth of research has shown that many GPCRs have much more complex signalling behaviour. For example, β_2AR exhibits significant constitutive activity, which can be blocked by inverse agonists[3,4]. The β_2AR couples to both Gαs and the inhibitory subunit (Gαi) in cardiac myocytes[5], and can also signal through MAP kinase pathways in a G-protein-independent manner through arrestin[6,7]. Similarly, the process of GPCR desensitization involves multiple pathways, including receptor phosphorylation events, arrestin-mediated internalization into endosomes, receptor recycling and lysosomal degradation[8,9]. These activities are further complicated by factors such as GPCR oligomerization[10], localization to specific membrane compartments[11] and resulting differences in lipid-bilayer composition. Such multifaceted functional behaviour has been observed for many different GPCRs.

How does this complex functional behavior reconcile with the biochemical and biophysical properties of GPCRs? The effect of a ligand on the structure and biophysical properties of a receptor, and hence on the biological response, is known as the ligand efficacy. Natural and synthetic ligands can be grouped into different efficacy classes (Fig. 1, inset): full agonists are capable of maximal receptor stimulation; partial agonists are unable to elicit full activity even at saturating concentrations; neutral antagonists have no effect on signalling activity but can prevent other ligands from binding to the receptor; and

[1]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, 279 Campus Drive, Palo Alto, California 94305, USA.

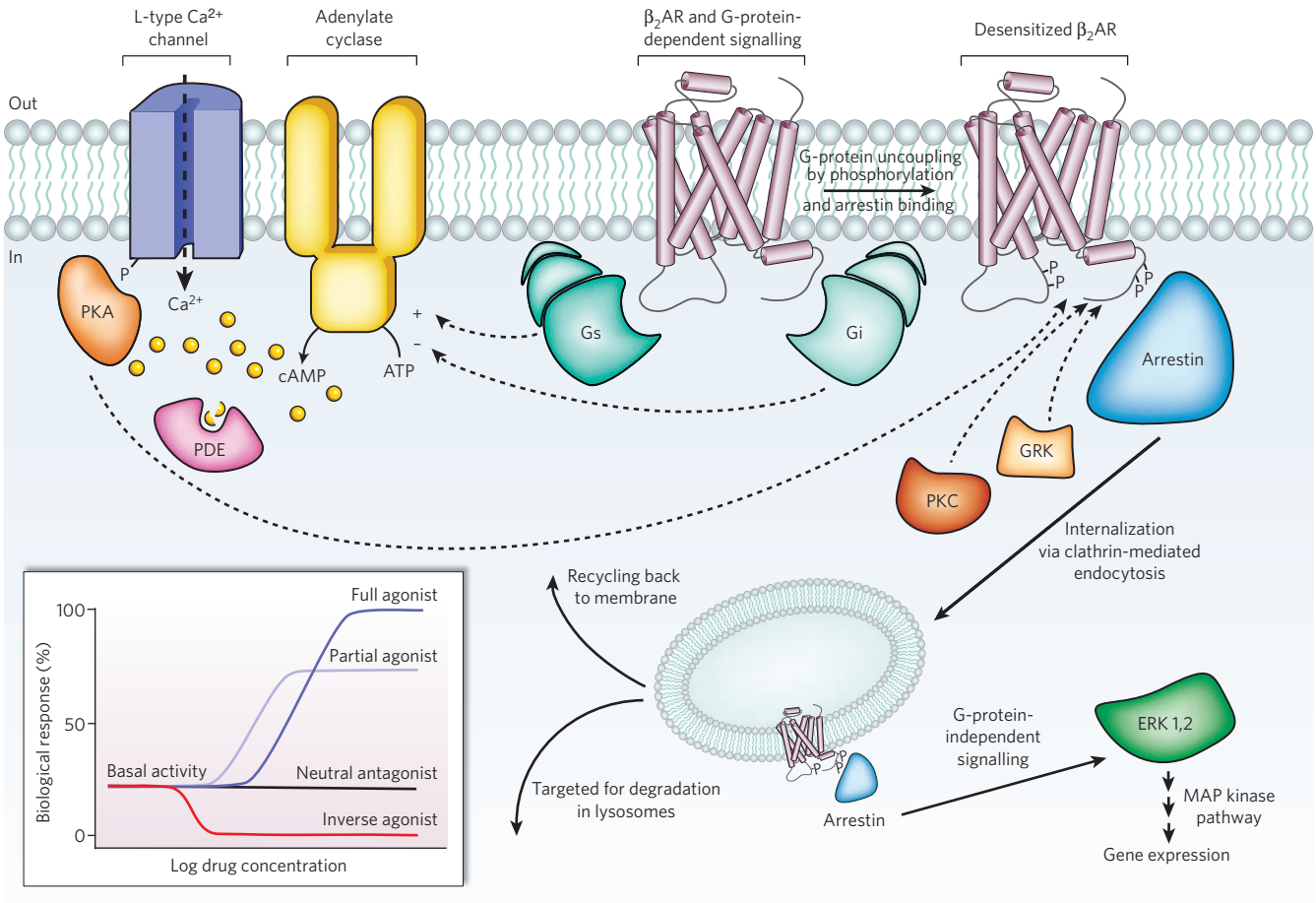


Figure 1 | Signal transduction in G-protein-coupled receptors. Diverse signalling pathways regulated by the type 2 beta adrenergic receptor (β_2AR). The β_2AR can activate two G proteins, Gαs and Gαi (part of the Gs and Gi heterotrimers, respectively), which differentially regulate adenylate cyclase. Adenylate cyclase generates cyclic AMP (cAMP), which activates protein kinase A (PKA), a kinase that regulates the activity of several cellular proteins including the L-type Ca^{2+} channel and the β_2AR. cAMP second messenger levels are downregulated by specific phosphodiesterase proteins (PDEs). Activation of the β_2AR also leads to phosphorylation by a G-protein-coupled receptor kinase (GRK) and subsequent coupling to arrestin. Arrestin is a signalling and regulatory protein that promotes the activation of extracellular signal-regulated kinases (ERK), prevents the activation of G proteins and promotes the internalization of the receptor through clathrin-coated pits. PKC, protein kinase C. The inset shows classification of ligand efficacy for GPCRs. Many GPCRs exhibit basal, agonist-independent activity. Inverse agonists inhibit this activity, and neutral antagonists have no effect. Agonists and partial agonists stimulate biological responses above the basal activity. Efficacy is not directly related to affinity; for example, a partial agonist can have a higher affinity for a GPCR than a full agonist.

inverse agonists reduce the level of basal or constitutive activity below that of the unliganded receptor. The wide spectrum of ligand efficacies for individual GPCRs shows that efficient energy transfer between the binding pocket and the site of G-protein interaction is dependent on multiple interactions between receptor and hormone, and requires more than simply occupying the binding site. Further, biophysical studies on purified fluorescently labelled β_2AR demonstrated that partial and full agonists containing different subsets of functional groups stabilize distinct conformational states by engaging with distinct subsets of conformational switches in the receptor[12–14]. These findings lead to a complex picture of GPCR activation in which a distinct conformation stabilized by a ligand's structure determines the efficacy towards a specific pathway. Many GPCRs can stimulate multiple signalling systems, and specific ligands can have different relative efficacies to different pathways[15]. In the extreme case, even opposite activities for different signalling pathways are observed: for β_2AR, agonists for the arrestin/MAP kinase pathway are also inverse agonists for the classical Gαs/cAMP/PKA pathway[7,16]. GPCRs are no longer thought to behave as simple two-state switches. Rather, they are more like molecular rheostats, able to sample a continuum of conformations with relatively closely spaced energies[17]. Specific ligands achieve varying efficacies for different signalling pathways by stabilizing particular sets of conformations that can interact with specific effectors.

The inactive states of four GPCRs

The first insights into the structure of GPCRs came from two-dimensional crystals of rhodopsin[18,19]. These structures revealed the general architecture of the seven transmembrane helices. However, given the conformational complexity of ligand-activated GPCRs, it is not surprising that it took so long to obtain three-dimensional crystal structures. As detailed in Box 1, a variety of different protein-modification and engineering approaches have contributed to recent advances in GPCR crystallography. We now have inactive-state structures of four GPCRs for comparison: human β_2AR bound to the high-affinity inverse agonists carazolol[20–22] and timolol[23]; avian β_1AR bound to the antagonist cyanopindolol[24]; the human A_{2A} adenosine receptor bound to the antagonist ZM241385 (ref. 25); and bovine rhodopsin[26–28] containing the covalently bound inverse agonist 11-*cis* retinal. The superpositions of different receptors using the homologous transmembrane domains led to root mean squared deviation (r.m.s.d.) values of less than 3 Å. This degree of overlap indicates that these four proteins have a similar overall architecture, yet the divergences are still high enough to signify important differences in helical packing interactions (Fig. 2).

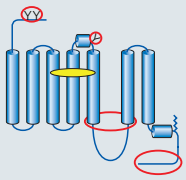
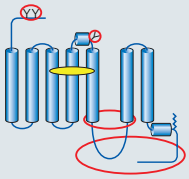
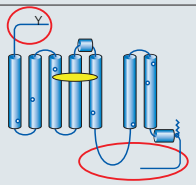
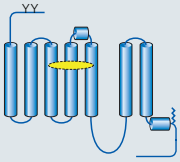
Extracellular surfaces and ligand-binding sites

As might be expected from the functional differences between the receptors, the most significant structural divergences lie in the extracellular

loops and ligand-binding region (Fig. 2b). The second extracellular loop (ECL2) of rhodopsin forms a short β-sheet that caps the covalently bound 11-*cis* retinal, shielding the chromophore from bulk solvent and preventing Schiff base hydrolysis. Further, the glycosylated amino terminus of rhodopsin adopts a structured conformation at the extracellular apex of the protein that further shields the covalently bound ligand[27,28]. In contrast, the ECL2 regions of β_1AR and β_2AR contain a short α-helix that is stabilized by intra- and inter-loop disulphide bonds, and the cytoplasmic N-terminal regions are disordered[21,22,24]. oIn the A_{2A} receptor, the ECL2 region lacks a predominant secondary structure, although the loop has multiple disulphide bonds that constrain the observed conformation and expose the ligand-binding cavity to extracellular bulk solvent[25]. Additional structures will be needed to confirm whether the open binding pocket is a general feature of GPCRs

Box 1 | Challenges in GPCR crystallography

Summary of structural modifications required to obtain crystals of G-protein-coupled receptors (GPCRs)

GPCR	Topology	Ligand (indicated by yellow oval)	Method of stabilization	Other modifications (indicated by red ovals and circles)
β_2AR-Fab		Inverse agonist for stabilization	Stabilized TM5/TM6 (transmembrane) region through binding to Fab, which improves crystal lattice-forming contacts	Truncated flexible C terminus (potential Ser/Thr phosphorylation sites) Removed N-linked glycosylation N-terminal affinity tag
β_2AR-T4 lysozyme Adenosine A_{2A}-T4 lysozyme		Inverse agonist or antagonist for stabilization	Stabilized TM5/TM6 region through T4L insertion, which improves crystal lattice-forming contacts	Truncated flexible ICL3 (intracellular loop) and C terminus (potential Ser/Thr phosphorylation sites) Removed N-linked glycosylation N- or C- terminal affinity tag
β_1AR		Antagonist for stabilization	Mutations to increase thermal stability and functional expression (indicated by blue circles)	Truncated flexible ICL3, N- and C-termini (glycosylation and Ser/Thr phosphorylation sites) Removed palmitoylation site C-terminal affinity tag
Rhodopsin/opsin		Crystal structures obtained with and without bound retinal	Native protein crystallizable without modifications	Not required

The first major challenge in GPCR crystallography is that most GPCRs are expressed at low levels in native tissues. A suitable recombinant expression system must therefore be developed to generate natively folded membrane protein. So far only Sf9 and Hi5 insect cells and COS-1 mammalian cells[64] have produced enough purified GPCR for structure determination (for bovine rhodopsin, a high level of expression in native rod-cell disc membranes allows purification from a natural source[65]).

The second major challenge is overcoming thermodynamic and proteolytic protein-stability problems (see the table above). GPCRs other than rhodopsin typically have poor thermal stability[66] and are prone to proteolysis as a result of their disordered extramembranous loop regions. In recent successful structural efforts, methods used to enhance the thermal stability of the target GPCRs include stabilizing ligands[20–22,25] (β_2AR, β_2AR-T4L and A_{2A}-T4L), a combination of stabilizing mutations[24,66] (β_1AR), the addition of lipids during purification and crystallization[20–22,25] (β_2AR, β_2AR-T4L and A_{2A}-T4L), and having a high salt concentration[25] (A_{2A}-T4L). Methods to enhance proteolytic stability include the truncation of disordered regions[20–22,25] (β_2AR, β_2AR-T4L, β_1AR and A_{2A}-T4L), fusion of a stable, well-folded protein domain at the third intracellular loop[21,25] (β_2AR-T4L and A_{2A}-T4L) and complex formation with an antibody Fab fragment[20,67] (β_2AR). Structural studies on dark-state bovine rhodopsin did not require such modifications, but purification and crystallization of the protein from a recombinant source did benefit from the engineering of a stabilizing disulphide bond between extracellular loop regions[64]. Further, the crystals of opsin[44,45] were formed at low pH, which is known to stabilize an active conformation of the retinal-free receptor[47].

Beyond the purification of large quantities of homogeneous stable membrane protein, several recent successful structural efforts relied on modifications to coax GPCRs into crystals (see the table above). For detergent-solubilized membrane proteins in general, and GPCRs specifically, the absence of significant exposed polar surface area outside the micelle can be a major impediment to crystallization. This was a crucial motivation for both the antibody Fab complex approach[20,67] and the T4L fusion strategy[21] for the β_2AR. In both these structures, as well as in the subsequent A_{2A}-T4L structure, most of the lattice contacts stabilizing the crystals involved the bound antibody or the fused T4L domain. It remains to be seen how general these approaches are to other GPCRs, although the application of the T4L strategy to the A_{2A} receptor[25] is a promising sign that other similarly engineered receptor structures may follow. Finally, the β_2AR/Fab complex and both T4L fusion-protein structures relied on lipid-mediated crystallogenesis, the former based on the bicelle methodology[68,69] and the latter based on lipidic mesophase techniques[70]. Although the examples of rhodopsin and the mutation-stabilized β_1AR show that these strategies are not absolutely necessary for crystal formation and X-ray structure determination of GPCRs, it is likely that most native receptors will not succumb easily to traditional methods of membrane-protein structural biology. The alternative engineering strategies described above are not without risk, namely that the introduction of modifications will alter or skew certain native features of the receptor. An example is seen in the β_2AR-T4L structure[21,22], where an arginine residue from T4 lysozyme forms a salt bridge with the mechanistically important Glu 268 from the receptor. Therefore, when these alternative strategies are applied to other GPCRs, it is important to rigorously characterize the modified proteins for native-like pharmacological and biophysical properties.

that recognize diffusible small-molecule ligands.

How similar are the ligand-binding pockets of these four receptors? The sites of carazolol, cyanopindolol and 11-*cis* retinal binding are partly overlapping in superpositions of β_2AR, β_1AR and rhodopsin, respectively (Fig. 3a). However, the overall positions of the ligands within the β_2AR and β_1AR structures are slightly more extracellular than 11-*cis*-retinal in rhodopsin. This difference in the positions of the ligands results in a significant difference in inverse-agonist–antagonist interactions with the residues Trp 286$^{6.48}$ (human β_2AR, with Ballesteros/Weinstein numbering[29] in superscript) and Trp 303$^{6.48}$ (avian β_1AR), which are suggested to undergo key rotamer conformational transitions in GPCR activation, referred to as the 'rotamer toggle switch'[30].

Partial agonism of β_2AR can be achieved by ligands without engaging the toggle switch, but full agonism appears to require this conformational change[14]. The ionone ring of retinal makes direct contact with the analogous Trp residue in rhodopsin, whereas carazolol in β_2AR and cyanopindolol in β_1AR pack against aromatic residues that shield the residue from the binding site. The less direct link between inverse-agonist binding and the inactive conformation of the rotamer toggle switch in β_2AR may help to explain the elevated basal activity of this receptor relative to rhodopsin. For both adrenergic receptors and rhodopsin, ligand binding is mediated by polar and hydrophobic contact residues from transmembrane helices 3, 5, 6 and 7 (TM3, TM5, TM6 and TM7). In contrast to the β_2AR, β_1AR and rhodopsin structures, the ligand ZM241385 binds to the A_{2A} receptor in a mode that is roughly perpendicular to the bilayer plane, and the packing interactions with the protein, mostly with TM6 and TM7, extend all the way from the toggle switch Trp 246$^{6.48}$ to the extracellular loops[25]. This comparison shows that, despite the highly conserved seven-transmembrane architecture, GPCRs can support a wide variety of ligand-binding modes that have differing degrees of interaction with regions involved in known conformational switches.

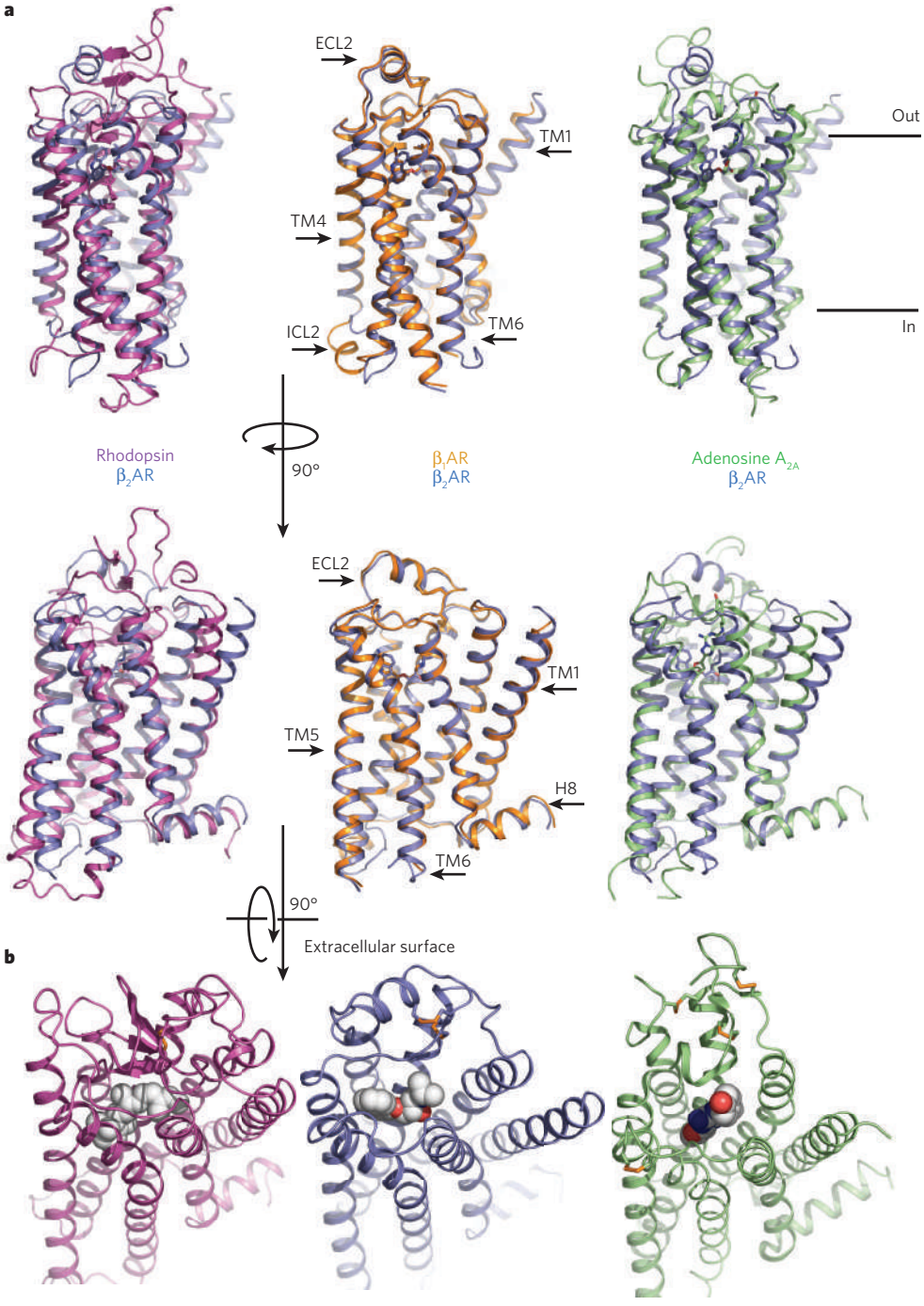


Figure 2 | Comparison of four GPCR structures. a, Bovine rhodopsin (purple), avian β_1AR (orange) and human A_{2A} adenosine receptor (green) are each superimposed on the human β_2AR structure (blue). The extracellular loop 2 (ECL2), intracellular loop 2 (ICL2), cytoplasmic helix 8 (H8) and several of the transmembrane segments are indicated on one of the structures. The greatest diversity in these structures lies in the extracellular ends of the transmembrane helices and the connecting loops. **b**, Extracellular views of rhodopsin, the β_2AR and the A_{2A} adenosine receptor. The ligands are shown as spheres.

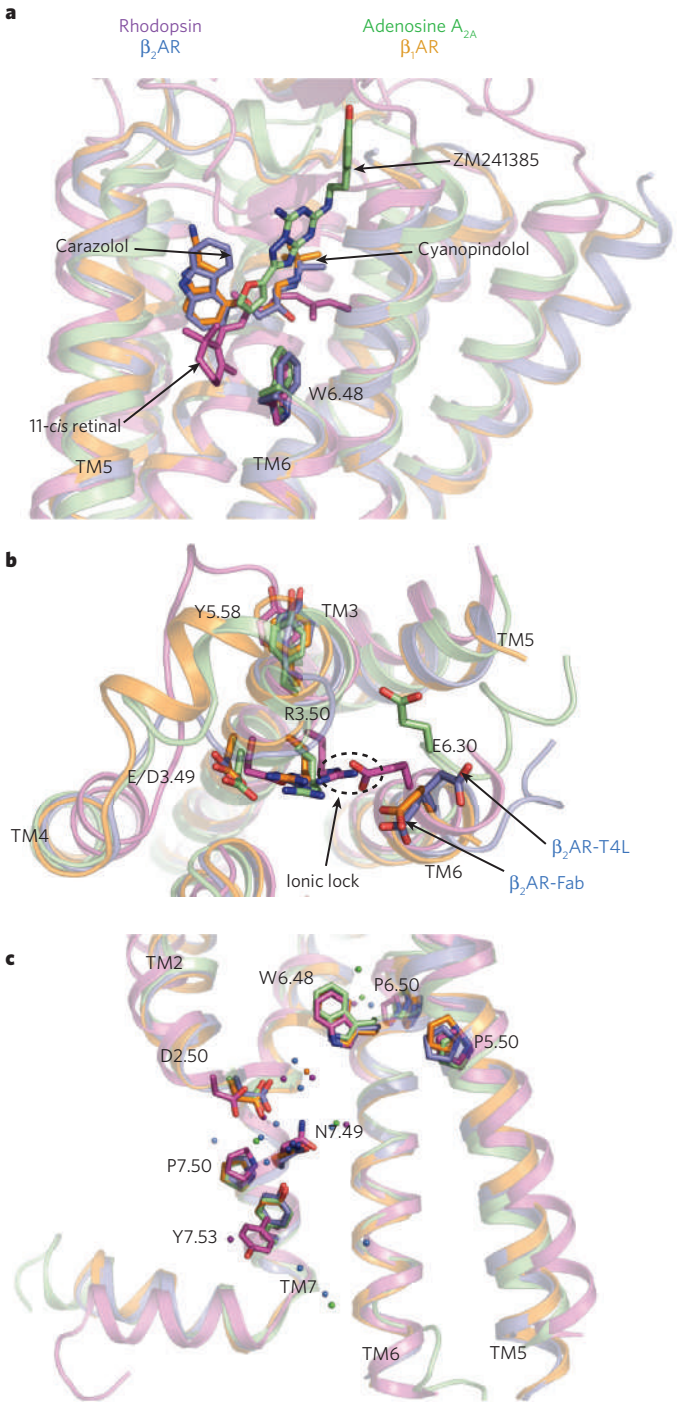


Figure 3 | Comparison of conserved regions of four GPCR structures. a, The locations of bound ligands for the four superimposed receptor structures bovine rhodopsin (purple, bound to 11-*cis* retinal), avian β_1AR (orange, bound to cyanopindolol), human A_{2A} adenosine receptor (green, bound to ZM241385) and human β_2AR (blue, bound to carazolol) are shown. W6.48 is the key residue of the rotamer toggle switch. TM, transmembrane segment. **b**, The ionic-lock residues at the cytoplasmic end of TM3 (R3.50 and E/D3.49), and TM6 (E6.30) are shown for the same four structures. R3.50 engages Y5.58 on TM5, rather than E6.30 on TM6 in the opsin 'active state'. The rotameric position of E6.30 differs for the two β_2AR structures. **c**, The location of several highly conserved residues around a cluster of water molecules (coloured spheres) is shown. These residues may be part of a common pathway for propagating conformational changes from the ligand-binding pocket to the G-protein coupling domains. Amino acids are numbered using the Ballesteros/Weinstein numbering system[29], in which the number preceding the dot refers to the transmembrane helix on which an amino acid resides. The second number designates the position relative to the most highly conserved residue among family A GPCRs, numbered 50.

The structures of avian β_1AR bound to cyanopindolol[24] and human β_2AR bound to carazolol[21,22] have almost identical binding pockets. This finding is expected owing to the related function of the proteins and the almost complete conservation of binding-site contact residues[31]. This high conservation in the ligand-binding pocket is also observed in other subfamilies of GPCRs (such as dopamine, serotonin and histamine), and probably explains some of the difficulty in obtaining potent subtype-selective compounds in pharmaceutical discovery programs[32]. Nevertheless, subtype-specific binding affinities are observed for β_1AR and β_2AR[33,34]. These differences cannot be based primarily on the amino acids forming the binding pocket, but involve more subtle influences on the arrangement of these amino acids as a result of subtype-specific conformational preferences in more distant residues.

Cytoplasmic surfaces of the GPCR structures

At the cytoplasmic surface, a major structural difference between the ligand-activated GPCRs and rhodopsin lies in the 'ionic lock' between the highly conserved E/DRY motif on TM3 and a glutamate residue on TM6 (Fig. 3b). Conserved among all family A GPCRs, these amino acids form a network of polar interactions that bridges the two transmembrane helices, stabilizing the inactive-state conformation[35]. For β_2AR, mutation of these residues increases constitutive activity[36,37], and biophysical studies have shown that both full and partial agonists can modulate the structure around the ionic lock[38]. This interaction network has been observed in dark-state rhodopsin crystals[27,28], but the analogous polar interactions are broken in all the ligand-activated GPCR structures, β_2AR, β_1AR and A_{2A}. The lack of an intact ionic lock in the crystal structures can be interpreted in two ways: either this interaction does not exist in the captured ligand-bound states, or the interaction is so weak that it is overcome energetically by various crystal packing forces. This observation is compatible with the findings that ligand-activated GPCRs generally have higher basal activity than rhodopsin[39]. Polar contact between adjacent acidic and basic residues on TM3 ($E/D^{3.49}$ followed by $R^{3.50}$ of the E/DRY motif; Fig. 3b) is maintained in all four inactive-state structures, and this interaction is also likely to be inhibitory to conformational changes leading to the active state (see below).

One common feature at the cytoplasmic surface of the structures is the chemical environment surrounding residues of the highly conserved NPXXY motif[40] (Fig. 3c). The cytoplasmic end of TM7, in which this motif is located, participates in key conformational changes associated with GPCR activation (see below). In all the structures, the proline in this motif causes a distortion in the α-helical structure, and the tyrosine faces into a pocket lined by TM2, TM3, TM6 and TM7. First observed in rhodopsin[28,41] and later seen in the high-resolution structures of β_2AR and A_{2A}, networks of ordered water molecules in this region help to reinforce the helical deformation of TM7 and provide hydrogen-bonding partners to polar side chains. Although this 'water pocket' network is presumed to stabilize the inactive state, the relative ease of breaking these weakly favourable solvent-mediated interactions probably allows for rapid toggling to the active state when an agonist binds[28,41,42].

A final difference between the structures at the cytoplasmic surface is found in the intracellular loop 2 region (ICL2), which includes a short α-helix in the β_1AR and A_{2A} structures that is absent in β_2AR and rhodopsin. This structure serves as a platform for a hydrogen-bonding interaction of a conserved tyrosine (on ICL2) with the E/DRY motif (on TM3); its absence in β_2AR could help to explain the higher relative basal activity of this receptor. Despite their differences, the four inactive-state GPCR structures are in close agreement. Figure 3c shows several of the most highly conserved residues in family A GPCRs, mapped onto the superimposed structures of β_2AR, β_1AR, A_{2A} and rhodopsin (including the rotamer toggle switch tryptophan, the NPXXY motif and several prolines that induce structurally important helical deformations). The clustering of these residues in the cytoplasmic half of the transmembrane bundle reflects the basic conservation of mechanism across GPCRs, a remarkable structural affirmation of a hypothesis made more than 20 years ago[43].

Active state of a GPCR in opsin crystals

The fundamental question of the mechanism for ligand-activated GPCRs remains: how does binding of an agonist, and the resulting changes in interactions at the ligand-binding pocket, lead to conformational changes that are propagated from the extracellular portion of the molecule to the cytoplasmic surface involved in G-protein binding. The recent structures of opsin provide clues to the transmembrane helix rearrangements that can be expected as a result of agonist binding[44,45]. Opsin is the retinal-free photoreceptor protein generated after photoactivation and Schiff base hydrolysis of rhodopsin. After photobleaching, rod photoreceptors exhibit residual activity that is presumed to result from basal activity of the unliganded state of rhodopsin[46]. On the basis of biochemical and infrared spectroscopic characterization, opsin at low pH is thought to be stabilized in an active state that resembles metarhodopsin II[47,48].

In the crystal structure of opsin at low pH[44], there are several subtle changes in the conformations of binding-pocket residues, relative to rhodopsin. Most importantly, the side chain of Trp $265^{6.48}$ (the toggle switch) moves into space previously occupied by the ionone ring of retinal, and there is only weak electron density for the Schiff base-forming Lys $296^{7.43}$ (on TM7). The interaction between Lys $296^{7.43}$ and the Schiff base counterion Glu $113^{3.28}$ (on TM3) is broken, and the pocket becomes slightly wider than in rhodopsin. Recent solid-state nuclear magnetic resonance (NMR) studies provide evidence for conformational changes that disrupt a hydrogen-bond network between ECL2 and the extracellular ends of TM4, TM5 and TM6 in metarhodopsin II before the dissociation of retinal and the formation of opsin[49].

More dramatic structural changes are observed at the cytoplasmic surface of the molecule. The cytoplasmic end of TM6 is shifted more than 6 Å outwards from the centre of the bundle relative to its position in the inactive state, and at the same time moves closer to TM5 (Fig. 4a, b). This rigid-body movement is consistent with previous biophysical studies of both rhodopsin[50,51] and β_2AR[38]. The new position of the cytoplasmic end of TM6 is stabilized by changes in several key interactions (Fig. 4b). Most importantly, the ionic lock is broken and new interactions are formed between Arg $135^{3.50}$ (of the ERY motif on TM3) and Tyr $223^{5.58}$ (TM5), as well as between Glu $247^{6.30}$ (TM6) and Lys $231^{5.66}$ (TM5) (Fig. 4b). This rearrangement and engagement of the ionic-lock residues in new interactions is distinct from the merely broken state of the ionic lock seen in the ligand-activated GPCRs. Additionally, Tyr $306^{7.53}$ from the NPXXY motif on TM7 undergoes a conformational change and inserts into space occupied by TM6 in dark-state rhodopsin, stabilizing the active conformation. The end result of the changes from inactive rhodopsin to active-state opsin is the creation of a cavity between TM3, TM5 and TM6 in which the G protein transducin can bind (Fig. 4c).

The structure of opsin bound to a carboxy-terminal peptide of transducin demonstrates that this cleft on the receptor does indeed provide the interaction surface for the most crucial binding epitope of the G protein[45]. Here the rearranged ionic-lock residues prove critical for the formation of the receptor–transducin peptide complex, notably where Arg $135^{3.50}$ of the ERY motif dissociates from Glu $134^{3.49}$ and forms the base of the peptide-binding cavity with stabilizing contacts from Tyr $223^{5.58}$ on TM5. The transducin-derived peptide adopts a C-capped α-helical structure and interacts with the receptor in an amphipathic manner: hydrophobic residues on one face of the transducin helix bind to a hydrophobic surface at the cytoplasmic ends of TM5 and TM6. The orientation of binding is enforced by a hydrogen-bonding network between the transducin C-cap and TM3, TM5 and helix 8 of opsin.

Considering the conserved three-dimensional structure and G-protein signalling mechanism between family A (rhodopsin family) GPCRs, it is reasonable to suppose that the activation of other GPCRs by diffusible ligands will be accompanied by similar changes in transmembrane helix packing to those observed in the opsin structures. In fact, biophysical studies of β_2AR are in good agreement with such a mechanism[38]. However, the question of how agonist binding far from the cytoplasmic surface leads to the expected packing rearrangements remains unanswered. In the β_2AR–carazolol and β_1AR–cyanopindolol

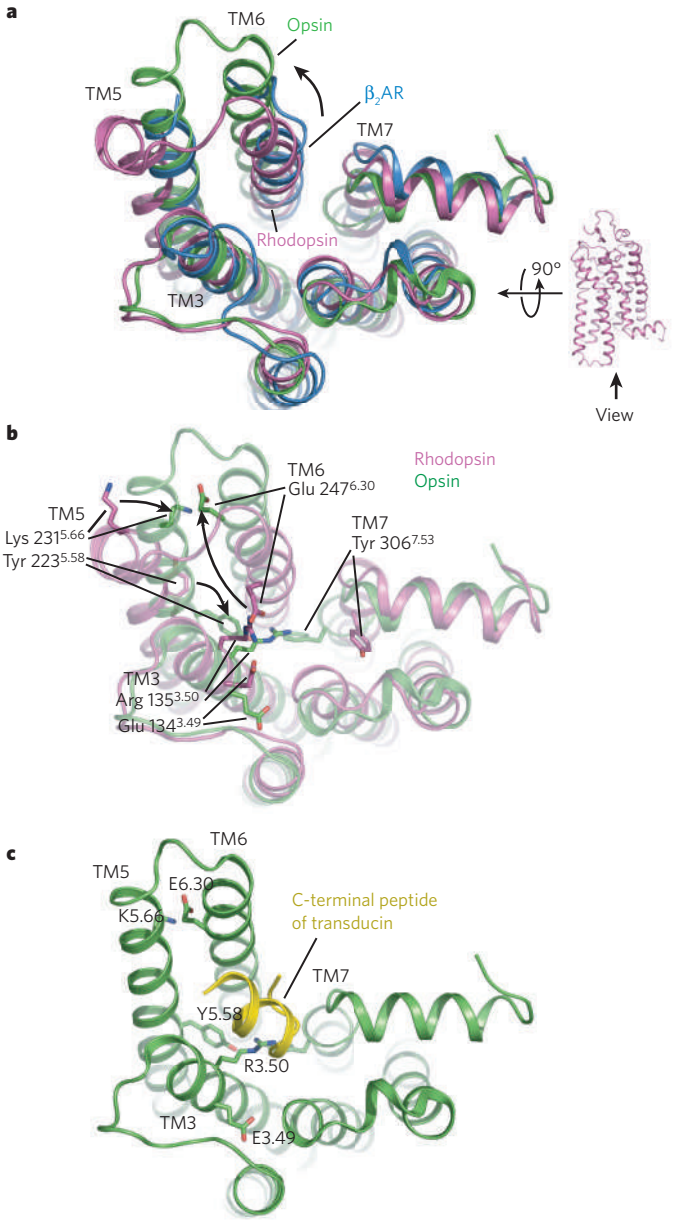


Figure 4 | The structure of opsin obtained at low pH represents an active form of rhodopsin. **a**, A comparison of the cytoplasmic surface of β_2AR (blue), rhodopsin (purple) and opsin (green). With the exception of transmembrane segment 5 (TM5), β_2AR is more similar to rhodopsin than to opsin. **b**, Differences between rhosopsin and opsin in interactions between conserved amino acids, including those of the ionic lock. **c**, Complex between opsin and a peptide representing the carboxyl terminus of the G protein transducin.

complexes, the captured inactive-state conformations cannot allow for simultaneous contacts between known agonist-binding amino acids and both ends of the catecholamine scaffold[21,24]. The incompatibility between the inactive-state adrenergic-receptor structures and agonist binding is analogous to the fact that the retinal binding pocket in the dark state of rhodopsin cannot accommodate the photon-activated all-*trans* conformation of the chromophore[27,28].

Using β_2AR and β_1AR as models, we conclude that conformational changes at the ligand-binding site must accompany agonist binding. One hypothesis is that the upper region of TM5, which contains several catechol-binding serines[52,53], moves closer to TM3. Simultaneous engagement of the agonist by TM5–catechol hydrogen-bonding and TM3/TM7–amine polar contacts (also essential for agonist binding[54])

would facilitate changes in the packing of nearby aromatic amino acids that shield Trp 268[6.48]. In this manner, the binding of an agonist could be coupled to movements of the rotamer toggle switch. The resulting conformational change could then lead to tightly coupled packing rearrangements that propagate towards the cytoplasmic surface. This hypothesis is supported by the central and buried locations of residues in the adrenergic receptors whose mutation confers constitutive activity[21] ('CAM mutants'); disruption of these packing interactions would allow freer transmembrane helix motions in the absence of an agonist. In the case of the A_{2A} receptor, the rotamer toggle switch is partly exposed at the base of the ligand-binding pocket and barely interacts with the buried furan ring of bound ZM241385 (ref. 25). The antagonist mainly contacts amino acids on TM5, TM6 and TM7, but the position on the adenine ring of ZM241385 to which the ribose moiety would be attached in the natural agonist adenosine orients the sugar towards TM3. Mutations at this precise region of TM3, analogous to an essential position for agonist-binding in β_2AR and β_1AR, have been shown to decrease agonist affinity to the A_{2A} receptor[55]. Overall, it is less apparent for the A_{2A} receptor how agonists might change the structure of the binding cavity, as seen in the inactive-state crystal structure. However, we can speculate that agonists with the ribose functional group would promote the engagement of TM3 residues, resulting in small changes in the relative transmembrane helix dispositions that could activate the rotamer toggle switch.

Future directions

What are the applications of these new GPCR structures, and what are the goals for future investigations? First, there is great interest in structural information to help guide GPCR drug discovery. Until recently, pictures of three-dimensional drug–receptor interactions could only be provided through speculative homology models based on rhodopsin[56]. For the GPCRs whose structures have now been solved, these modelling efforts have been shown to be imprecise at the level required by *in silico* drug designers. With the inactive-state structures of β_2AR, β_1AR and the A_{2A} receptor, pharmaceutical chemists now have experimental data to guide the development of ligands for several active therapeutic targets. However, the value of these high-resolution structures for *in silico* screening may be limited. Recent molecular docking studies using the β_2AR crystal structure as a template identified six new β_2AR ligands that bound with affinities ranging from 9 nM to 4 μM; however, every compound exhibited inverse agonist activity. These results suggest that structures of inactive GPCRs will only be reliable for identifying compounds that stabilize the inactive state[57].

In a broader sense, the success of these efforts proves that obtaining the crystal structures of GPCR–drug complexes, although still extremely challenging, is at least tractable. Nevertheless, the structures available represent only a small proportion of GPCRs, as implied by their relatively close phylogenetic relationships[1]. There are still no crystal structures for most of the main branches of the rhodopsin family, or for other GPCR families with large differences in architecture, such as the GABA or (γ-aminobutyric acid) mGluR receptors in family C. Validated drug targets are present throughout the GPCR phylogeny, making it vitally important to develop crystallization methods that can be applied to receptors distantly related to rhodopsin and the biogenic amines. The high-resolution crystallography of GPCRs will hopefully become as routine a tool for drug development as that of kinases.

Beyond the crystallization of more GPCRs, we must develop methods for acquiring structures of receptors bound to agonists. The opsin crystals, without bound retinal but prepared under low-pH activating conditions, have provided a molecular picture of a state resembling fully active metarhodopsin II. Similarly, agonist-bound receptor crystals may provide three-dimensional representations of the active states of other GPCRs. These structures will help clarify the conformational changes connecting the ligand-binding and G-protein-interaction sites, and lead to more precise mechanistic hypotheses. GPCR-targeted therapeutics include agonists as well as antagonists, so these structures will have a broader impact extending to medicinal chemistry and pharmacology. Given the conformational flexibility inherent to ligand-activated GPCRs and the greater heterogeneity exhibited by agonist-bound receptors[58], stabilizing such a state will not be easy. The crystal structure of a photoactivated deprotonated intermediate of rhodopsin[59] illustrates that a G-protein-interacting state of a GPCR may not be captured in a given crystal lattice, even with a covalent full agonist occupying the binding pocket. Indeed, the possibility of a deprotonated intermediate of rhodopsin in an inactive conformation was directly demonstrated by kinetic electron paramagnetic resonance (EPR) measurements[60]. Ultimately, the true active state of GPCRs will only be revealed through the co-crystallization of receptors with G proteins, which will also help to reveal how agonist binding is coupled to nucleotide exchange across the protein–protein interface. Such efforts will benefit from the predicted stabilization of a homogeneous agonist-bound receptor conformation in the ternary complex[61], as well as the addition of a large soluble protein to participate in crystal-lattice formation. However, the complex dependency of this interaction on experimental conditions makes it difficult to trap a stable GPCR-G protein complex.

As important as the recent structures have been for GPCR research, crystallography has major limitations for characterizing and understanding these physiologically important receptors. As discussed above, GPCRs are inherently flexible proteins that are able to exhibit a spectrum of conformations depending on such factors as the presence of a bound ligand, the lipid environment and the presence of interacting proteins. The conformational dynamics of GPCRs are of more than academic interest: the stabilization of receptor states is the key to modulating GPCR function. To study the relationships between conformational states and the rates of interconversion between them, we need solution-based or membrane-compatible biophysical tools that make direct measurements of the relative positions of different receptor residues on a timescale consistent with the molecular motions. So far, fluorescence spectroscopy and EPR techniques have allowed the study of conformational changes for β_2AR[17] and rhodopsin[62], respectively; however, the application of other methods, such as NMR spectroscopy, promises to greatly expand our knowledge of GPCR dynamics[49,63]. Important structural properties of GPCRs, such as oligomerization, are not effectively addressed by crystallographic structures, and biophysical techniques can potentially be harnessed to study these phenomena. Only a marriage of biophysical methods with high-resolution X-ray crystallography will provide a full structural understanding of GPCR function. ■

1. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* **63,** 1256–1272 (2003).
This paper provides a comprehensive analysis of sequence relationships between G-protein-coupled receptors in the human genome.
2. Hoffman, B. B. & Lefkowitz, R. J. Adrenergic receptors in the heart. *Annu. Rev. Physiol.* **44,** 475–484 (1982).
3. Samama, P., Pei, G., Costa, T., Cotecchia, S. & Lefkowitz, R. J. Negative antagonists promote an inactive conformation of the beta 2-adrenergic receptor. *Mol. Pharmacol.* **45,** 390–394 (1994).
4. Chidiac, P., Hebert, T. E., Valiquette, M., Dennis, M. & Bouvier, M. Inverse agonist activity of beta-adrenergic antagonists. *Mol. Pharmacol.* **45,** 490–499 (1994).
5. Xiao, R. P., Cheng, H., Zhou, Y. Y., Kuschel, M. & Lakatta, E. G. Recent advances in cardiac beta(2)-adrenergic signal transduction. *Circ. Res.* **85,** 1092–1100 (1999).
6. Shenoy, S. K. *et al.* Beta-arrestin-dependent, G protein-independent ERK1/2 activation by the beta2 adrenergic receptor. *J. Biol. Chem.* **281,** 1261–1273 (2006).
7. Azzi, M. *et al.* Beta-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc. Natl Acad. Sci. USA* **100,** 11406–11411 (2003).
8. Freedman, N. J. & Lefkowitz, R. J. Desensitization of G protein-coupled receptors. *Recent Prog. Horm. Res.* **51,** 319–351; discussion 352–353 (1996).
9. Hanyaloglu, A. C. & von Zastrow, M. Regulation of GPCRs by endocytic membrane trafficking and its potential implications. *Annu. Rev. Pharmacol. Toxicol.* **48,** 537–568 (2008).
10. Terrillon, S. & Bouvier, M. Roles of G-protein-coupled receptor dimerization. *EMBO Rep.* **5,** 30–34 (2004).
11. Insel, P. A. *et al.* Caveolae and lipid rafts: G protein-coupled receptor signaling microdomains in cardiac myocytes. *Ann. NY Acad. Sci.* **1047,** 166–172 (2005).
12. Ghanouni, P., Steenhuis, J. J., Farrens, D. L. & Kobilka, B. K. Agonist-induced conformational changes in the G-protein-coupling domain of the beta 2 adrenergic receptor. *Proc. Natl Acad. Sci. USA* **98,** 5997–6002 (2001).
13. Swaminath, G. *et al.* Sequential binding of agonists to the beta2 adrenoceptor. Kinetic evidence for intermediate conformational states. *J. Biol. Chem.* **279,** 686–691 (2004).
14. Swaminath, G. *et al.* Probing the beta2 adrenoceptor binding site with catechol reveals differences in binding and activation by agonists and partial agonists. *J. Biol. Chem.* **280,** 22165–22171 (2005).

15. Galandrin, S., Oligny-Longpre, G. & Bouvier, M. The evasive nature of drug efficacy: implications for drug discovery. *Trends Pharmacol. Sci.* **28,** 423–430 (2007).
16. Wisler, J. W. *et al.* A unique mechanism of beta-blocker action: carvedilol stimulates beta-arrestin signaling. *Proc. Natl Acad. Sci. USA* **104,** 16657–16662 (2007).
17. Kobilka, B. K. & Deupi, X. Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **28,** 397–406 (2007).
18. Krebs, A., Villa, C., Edwards, P. C. & Schertler, G. F. Characterisation of an improved two-dimensional p22121 crystal from bovine rhodopsin. *J. Mol. Biol.* **282,** 991–1003 (1998).
19. Schertler, G. F., Villa, C. & Henderson, R. Projection structure of rhodopsin. *Nature* **362,** 770–772 (1993).
This paper presents the first three-dimensional structure of a G-protein-coupled receptor using cryoelectron microscopy of two-dimensional crystals.
20. Rasmussen, S. G. *et al.* Crystal structure of the human β_2 adrenergic G-protein-coupled receptor. *Nature* **450,** 383–387 (2007).
This is the first reported three-dimensional crystal structure of a ligand-activated G-protein-coupled receptor.
21. Rosenbaum, D. M. *et al.* GPCR engineering yields high-resolution structural insights into β_2-adrenergic receptor function. *Science* **318,** 1266–1273 (2007).
22. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human β_2-adrenergic G-protein-coupled receptor. *Science* **318,** 1258–1265 (2007).
23. Hanson, M. A. *et al.* A specific cholesterol binding site is established by the 2.8 Å structure of the human β_2-adrenergic receptor. *Structure* **16,** 897–905 (2008).
24. Warne, T. *et al.* Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature* **454,** 486–491 (2008).
25. Jaakola, V. P. *et al.* The 2.6 angstrom crystal structure of a human A_{2A} adenosine receptor bound to an antagonist. *Science* **322,** 1211–1217 (2008).
26. Palczewski, K. *et al.* Crystal structure of rhodopsin: A G-protein-coupled receptor. *Science* **289,** 739–745 (2000).
This paper presents the first three-dimensional crystal structure of a G-protein-coupled receptor, the visual photoreceptor rhodopsin.
27. Okada, T. *et al.* The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J. Mol. Biol.* **342,** 571–583 (2004).
28. Li, J., Edwards, P. C., Burghammer, M., Villa, C. & Schertler, G. F. Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.* **343,** 1409–1438 (2004).
29. Ballesteros, J. A. & Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein coupled receptors. *Methods Neurosci.* **25,** 366–428 (1995).
30. Shi, L. *et al.* β_2 adrenergic receptor activation. Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. *J. Biol. Chem.* **277,** 40989–40996 (2002).
31. Horn, F. *et al.* GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* **31,** 294–297 (2003).
32. Conn, P. J., Christopoulos, A. & Lindsley, C. W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nature Rev. Drug Discov.* **8,** 41–54 (2009).
33. Baker, J. G. The selectivity of β-adrenoceptor antagonists at the human β1, β2 and β3 adrenoceptors. *Br. J. Pharmacol.* **144,** 317–322 (2005).
34. Sugimoto, Y. *et al.* β_1-selective agonist (-)-1-(3,4-dimethoxyphenetylamino)-3-(3,4-dihydroxy)-2-propanol [(-)-RO363] differentially interacts with key amino acids responsible for β_1-selective binding in resting and active states. *J. Pharmacol. Exp. Ther.* **301,** 51–58 (2002).
35. Vogel, R. *et al.* Functional role of the "ionic lock"—an interhelical hydrogen-bond network in family A heptahelical receptors. *J. Mol. Biol.* **380,** 648–655 (2008).
36. Ballesteros, J. A. *et al.* Activation of the β_2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276,** 29171–29177 (2001).
37. Rasmussen, S. G. *et al.* Mutation of a highly conserved aspartic acid in the β_2 adrenergic receptor: constitutive activation, structural instability, and conformational rearrangement of transmembrane segment 6. *Mol. Pharmacol.* **56,** 175–184 (1999).
38. Yao, X. *et al.* Coupling ligand structure to specific conformational switches in the β_2-adrenoceptor. *Nature Chem. Biol.* **2,** 417–422 (2006).
39. Bond, R. A. & Ijzerman, A. P. Recent developments in constitutive receptor activity and inverse agonism, and their potential for GPCR drug discovery. *Trends Pharmacol. Sci.* **27,** 92–96 (2006).
40. Barak, L. S., Menard, L., Ferguson, S. S., Colapietro, A. M. & Caron, M. G. The conserved seven-transmembrane sequence NP(X)2,3Y of the G-protein-coupled receptor superfamily regulates multiple properties of the β_2-adrenergic receptor. *Biochemistry* **34,** 15407–15414 (1995).
41. Okada, T. *et al.* Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc. Natl Acad. Sci. USA* **99,** 5982–5987 (2002).
42. Pardo, L., Deupi, X., Dolker, N., Lopez-Rodriguez, M. L. & Campillo, M. The role of internal water molecules in the structure and function of the rhodopsin family of G protein-coupled receptors. *ChemBioChem* **8,** 19–24 (2007).
43. Dixon, R. A. *et al.* Cloning of the gene and cDNA for mammalian β-adrenergic receptor and homology with rhodopsin. *Nature* **321,** 75–79 (1986).
This paper reports the cloning of the first ligand-activated G-protein-coupled receptor.
44. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W. & Ernst, O. P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454,** 183–187 (2008).
45. Scheerer, P. *et al.* Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **455,** 497–502 (2008).
This paper presents the high-resolution stucture of an active-state G-protein-coupled receptor in complex with a G-protein peptide.
46. Lamb, T. D. & Pugh, E. N. Jr. Dark adaptation and the retinoid cycle of vision. *Prog. Retin. Eye Res.* **23,** 307–380 (2004).
47. Vogel, R. & Siebert, F. Conformations of the active and inactive states of opsin. *J. Biol. Chem.* **276,** 38487–38493 (2001).
48. Cohen, G. B., Oprian, D. D. & Robinson, P. R. Mechanism of activation and inactivation of opsin: role of Glu113 and Lys296. *Biochemistry* **31,** 12592–12601 (1992).
49. Ahuja, S. *et al.* Helix movement is coupled to displacement of the second extracellular loop in rhodopsin activation. *Nature Struct. Mol. Biol.* **16,** 168–175 (2009).
50. Farrens, D. L., Altenbach, C., Yang, K., Hubbell, W. L. & Khorana, H. G. Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* **274,** 768–770 (1996).
This is the first biophysical study to demonstrate movement of transmembrane segment 6 upon activation of rhodopsin.
51. Altenbach, C., Kusnetzow, A. K., Ernst, O. P., Hofmann, K. P. & Hubbell, W. L. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc. Natl Acad. Sci. USA* **105,** 7439–7444 (2008).
52. Strader, C. D., Candelore, M. R., Hill, W. S., Sigal, I. S. & Dixon, R. A. Identification of two serine residues involved in agonist activation of the β-adrenergic receptor. *J. Biol. Chem.* **264,** 13572–13578 (1989).
53. Liapakis, G. *et al.* The forgotten serine. A critical role for Ser-2035.42 in ligand binding to and activation of the β_2-adrenergic receptor. *J. Biol. Chem.* **275,** 37779–37788 (2000).
54. Strader, C. D. *et al.* Conserved aspartic acid residues 79 and 113 of the β-adrenergic receptor have different roles in receptor function. *J. Biol. Chem.* **263,** 10267–10271 (1988).
55. Jiang, Q., Lee, B. X., Glashofer, M., van Rhee, A. M. & Jacobson, K. A. Mutagenesis reveals structure-activity parallels between human A_{2A} adenosine receptors and biogenic amine G protein-coupled receptors. *J. Med. Chem.* **40,** 2588–2595 (1997).
56. Patny, A., Desai, P. V. & Avery, M. A. Homology modeling of G-protein-coupled receptors and implications in drug design. *Curr. Med. Chem.* **13,** 1667–1691 (2006).
57. Kolb, P. *et al.* Structure-based discovery of β_2-adrenergic receptor ligands. *Proc. Natl Acad. Sci. USA* **106,** 6843–6848 (2009).
58. Ghanouni, P. *et al.* Functionally different agonists induce distinct conformations in the G protein coupling domain of the β_2 adrenergic receptor. *J. Biol. Chem.* **276,** 24433–24436 (2001).
59. Salom, D. *et al.* Crystal structure of a photoactivated deprotonated intermediate of rhodopsin. *Proc. Natl Acad. Sci. USA* 103, 16123–16128 (2006).
60. Knierim, B., Hofmann, K. P., Ernst, O. P. & Hubbell, W. L. Sequence of late molecular events in the activation of rhodopsin. *Proc. Natl Acad. Sci. USA* **104,** 20290–20295 (2007).
61. De Lean, A., Stadel, J. M. & Lefkowitz, R. J. A ternary complex model explains the agonist-specific binding properties of the adenylate cyclase-coupled β-adrenergic receptor. *J. Biol. Chem.* **255,** 7108–7117 (1980).
62. Hubbell, W. L., Altenbach, C., Hubbell, C. M. & Khorana, H. G. Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site-directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Adv. Protein Chem.* **63,** 243–290 (2003).
63. Werner, K., Richter, C., Klein-Seetharaman, J. & Schwalbe, H. Isotope labeling of mammalian GPCRs in HEK293 cells and characterization of the C-terminus of bovine rhodopsin by high resolution liquid NMR spectroscopy. *J. Biomol. NMR* **40,** 49–53 (2008).
64. Standfuss, J. *et al.* Crystal structure of a thermally stable rhodopsin mutant. *J. Mol. Biol.* **372,** 1179–1188 (2007).
65. Okada, T. *et al.* X-ray diffraction analysis of three-dimensional crystals of bovine rhodopsin obtained from mixed micelles. *J. Struct. Biol.* **130,** 73–80 (2000).
66. Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. Conformational thermostabilization of the β_1-adrenergic receptor in a detergent-resistant form. *Proc. Natl Acad. Sci. USA* **105,** 877–882 (2008).
67. Day, P. W. *et al.* A monoclonal antibody for G protein-coupled receptor crystallography. *Nature Methods* **4,** 927–929 (2007).
68. Faham, S. & Bowie, J. U. Bicelle crystallization: a new method for crystallizing membrane proteins yields a monomeric bacteriorhodopsin structure. *J. Mol. Biol.* **316,** 1–6 (2002).
69. Faham, S. *et al.* Crystallization of bacteriorhodopsin from bicelle formulations at room temperature. *Protein Sci.* **14,** 836–840 (2005).
70. Caffrey, M. Crystallizing membrane proteins for structure determination: use of lipidic mesophases. *Annu. Rev. Biophys.* **38,** doi:10.1146/annurev.biophys.050708.133655 (2008).

Acknowledgements This work was supported by the US National Institute of General Medical Sciences (grant F32 GM082028 to D.M.R. and grant RO1-GM083118 to B. K.), the Lundbeck Foundation (to S.G.F.R.), the National Institute of Neurological Disorders and Stroke (grant RO1-NS28471 to B.K.) and the Mather Charitable Foundation (to B. K.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to B.K.K. (kobilka@stanford.edu).

NATURE|Vol 459|21 May 2009|doi:10.1038/nature08145

Torque generation and elastic power transmission in the rotary F_OF_1-ATPase

Wolfgang Junge[1], Hendrik Sielaff[1,2] & Siegfried Engelbrecht[1,3]

Adenosine triphosphate (ATP), the universal fuel of the cell, is synthesized from adenosine diphosphate (ADP) and inorganic phosphate (P_i) by 'ATP synthase' (F_OF_1-ATPase). During respiration or photosynthesis, an electrochemical potential difference of protons is set up across the respective membranes. This powers the enzyme's electrical rotary nanomotor (F_O), which drives the chemical nanomotor (F_1) by elastic mechanical-power transmission, producing ATP with high kinetic efficiency. Attempts to understand in detail the mechanisms of torque generation in this simple and robust system have been both aided and complicated by a wealth of sometimes conflicting data.

F_OF_1-ATPase uses transmembrane ion flow to drive the synthesis of ATP from ADP and phosphate. Because cells use ATP as their main energy supply, this enzyme is an essential machine of the power stations of the cell. In principle, it can operate in either direction, to synthesize ATP at the expense of ion flow, or to drive ion flow while hydrolysing ATP, although this sometimes occurs only in the forward direction. F_OF_1-ATPase is composed of two rotary motors coupled by a central rotary shaft and held together by an eccentric bearing. The coupled operation of two rotary motors, one electrical (F_O) and one chemical (F_1), is unique; most other motors are either linear (such as kinesin) or consist of a single rotary motor (examples range from helicases to the flagellar motor). F_OF_1-ATPase is well characterized as it has a simple construction and is easy to access experimentally. It has been used to study the mechanisms of torque (or force) generation in nanomotors and also the way that the stepped power strokes of a driving motor are connected to distinct molecular events in the objects it drives. In F_OF_1-ATPase there is no fine-tuning of the two stepping motors; instead, their coupled operation is smoothed and speeded by elastic power transmission, which accounts for its high kinetic efficiency and robust function. Other nanomotors probably share this feature.

In this Review, we discuss the proton-based driving force of ATP synthesis (Peter Mitchell's 'chemiosmotic mechanism'), the cooperativity between the chemical reaction sites on the F_1 motor (Paul Boyer's 'binding change mechanism'), and the stepping of rotation. We also propose a solution to the question of the nucleotide occupancy of F_1 and describe the orientation of the central shaft during the catalytic cycle. We then provide estimates of the magnitude of torque, its dependence on the angular position of the central shaft, and the elastic parameters of the enzyme, and we emphasize how the elastic power transmission between F_O and F_1 enhances their kinetic efficiency. Finally, we outline some open questions and challenges for the future.

Proton translocation and ATP synthesis

ATP is created mainly by photosynthesis and respiration. It was initially believed that the electron transport chains in these processes would directly generate a phosphorylated intermediate as a precursor to phosphoryl transfer to ADP. But in 1961, Mitchell realized that ATP formation might be powered by ion translocation[1], and ion-driven ATP synthesis has now been well documented in Archaea, Eubacteria, and the chloroplasts and mitochondria of Eukaryota.

The basic tenets of Mitchell's 'chemiosmotic hypothesis' are as follows. Metabolism drives vectorial electron transport, and thereby proton transfer across the membrane, which by itself is proton tight. The back-flow of protons caused by the protonmotive force is confined to the proton-translocating ATP synthase and coupled to the production of ATP. The driving force, termed 'protonmotive force', was originally conceived as the electrochemical potential difference of the proton between two aqueous 'bulk' compartments, each of which is in equilibrium with its membrane surface. This idea had to be modified in view of the real conditions of ATP synthesis in some systems. In the thylakoids of chloroplasts, the cristae of mitochondria and bacterial chromatophores, for example, the volumes can be so small that the number of freely diffusing protons is negligible; here protons are supplied to the enzyme by numerous proton-buffering groups at the membrane surface[2]. In alkaliphilic bacteria[3], which thrive at around pH 10, the bulk-to-bulk driving force at first sight seems insufficient. Under stationary operation, however, the pH close to the surface of the membrane and at moderate lateral distances between the pumps and synthases is more acidic than in the alkaline bulk phase, so the local protonmotive force between membrane surfaces is greater than anticipated, helping to drive ATP synthesis[4].

The search for a molecular mechanism started with the discovery in mitochondria of an ATPase[5]: namely, the water-soluble portion (termed the 'coupling factor', F_1) of a transmembrane protein, the membrane-embedded part of which is now known as F_O. Boyer and co-workers established that F_1 can hydrolyse ATP in at least two[6,7] equivalent reaction sites with alternating activity, by what he coined the 'binding change mechanism' (see ref. 8 and references therein), or in three such sites with rotary activity[9]. The first asymmetric crystal structure of bovine mitochondrial F_1 by John Walker and co-workers[10], resembling a three-chambered molecular Wankel engine, therefore strongly suggested that rotation, rather than alternation, was nature's choice, and that the synthesis of ATP might be mechanically driven by rotation of the central stalk in F_1. A rotary mechanism for ATP hydrolysis by F_1 was demonstrated by Hiroyuki Noji and Ryohei Yasuda in the laboratories of Masasuke Yoshida and Kazuhiko Kinosita by recording the rotation of the central shaft in

[1]Department of Biophysics, University of Osnabrück, Barbarastrasse 11, 49076 Osnabrück, Germany. [2]Present address: Institute of Biophysics, University of Ulm, 89081 Ulm, Germany. [3]Present address: Department of Biochemistry, University of Osnabrück, 49076 Osnabrück, Germany.

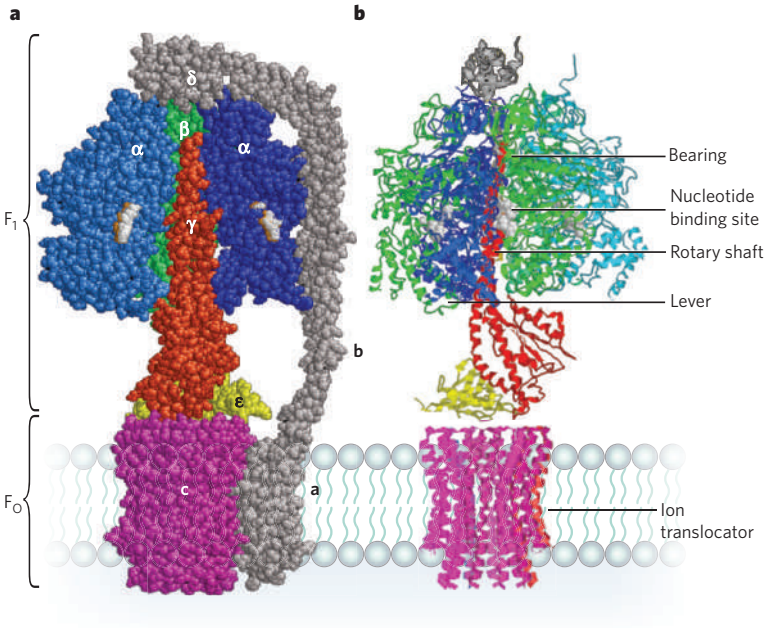


Figure 1 | Structural model of F_OF_1 in the bacterium *Escherichia coli*. **a**, Space-filling model with one α subunit and two β subunits removed to expose the γ subunit (red) in the centre of the $(\alpha\beta)_3$-pseudohexagon, with α shown in blues, β in green, γ in red, ε in yellow, δ (on top) and subunits a and b in dark grey, nucleotides in pale grey and c_{10} in magenta. The bulge of subunit γ (made up of the convex coiled-coil consisting of the N- and C-terminal helices of γ) is not visible as it points away from the viewer. **b**, Ribbon model lacking subunits a and b, and viewed from above, rotated by 90° anticlockwise from **a**. Labels point to the positions of the hydrophobic bearing of the central shaft, the nucleotide-binding sites (pale grey), the rotary shaft made up from subunit γ, and the lever formed by a loop of subunit β, which opens and closes the respective nucleotide-binding site when flexed by the rotation of subunit γ. The bulge of subunit γ points to the lever. Supplementary Movie 1 illustrates the action of subunit γ on the lever of subunit β (energy-minimized interpolation of the MF_1 crystal structure[10] by D. A. Cherepanov). The structural model is a composite of several partial structures (PDB codes 1BMF (ref. 10), 2CK3 (ref. 16), 1H8E (ref. 19) and 1YCE (ref. 52)) in the Protein Data Bank (http://www.pdb.org), here drawn by RasMol 2.7.2.1.1 (ref. 75), that formed the basis of a homology model of EF_OEF_1. The complete structures of δ, a and b are unknown. They have been modelled here in roughly their correct size and manually attached to the remainder, taking into account available crosslink data.

real time[11]). An idea for how proton transport might generate torque in F_O and drive the central stalk has complemented this view[12,13].

So, the ATP synthase F_OF_1 is composed of two rotary motors/generators that are mechanically coupled by a central rotor and an eccentric stator. The chemical function is confined to the soluble portion (F_1, in *Escherichia coli* subunits $\delta(\alpha\beta)_3\gamma\varepsilon$) and the ion-transporting function is limited to the membrane-intrinsic portion (F_O, subunits ab_2c_{10-15}) (Fig. 1). To account for their rotary mechanical coupling, the subunits are usually regrouped into the central rotor ($\gamma\varepsilon c_{10-15}$) and the stator ($ab_2\delta(\alpha\beta)_3$), although these are arbitrary assignments because the membrane-embedded enzyme as a whole is subject to rapid Brownian rotation[14].

F_OF_1 has a similar structure and operates according to the same principle in all photosynthetic and aerobic organisms, although there are species-dependent differences in subunit composition, preference of the electromotor for H^+ or Na^+, and redox and protonmotive-force regulation to meet the requirements of the cell (see ref. 15 for a review of this diversity).

F_1 is a rotary chemical motor and generator

Walker's group, studying bovine mitochondrial F_1, has revealed the asymmetric structure of the $(\alpha\beta)_3\gamma$ complex[10] (see also ref. 16 for a structure at 1.9 Å resolution). It has been interpreted as a still picture with three essentially equivalent catalytic sites, which, when driven by the rotation of subunit γ, change their role: one binds ADP and P_i, the next processes them into ATP, and the third releases ATP. This continues in a cycle, in line with biochemical concepts[8]. Figure 1 shows a homology model of *E. coli* F_OF_1 constructed from the adaptation of several partial structures obtained from different organisms, as no full structure of any particular F_OF_1 is yet available. The functional heart of F_1 is a pseudo-hexagon of subunits α and β, arranged as $(\alpha\beta)_3$. It carries a total of six nucleotide-binding sites[10], of which three are catalytic and three non-catalytic[17,18]. The catalytic sites are located mainly on β subunits at the interface with α. Subunit γ forms the central stalk of the hexagon, and the crystal structure suggests that γ rotates, with its bulge acting on a lever on subunit β such that the three catalytic sites are cycled through open and closed conformations for the uptake, processing and release of nucleotides, respectively (Supplementary Movie 1). The orientation of subunit γ dictates the momentary conformation of the three β subunits and hence their nucleotide occupancy.

Walker and colleagues have deposited an impressive total of 19 crystal structures of mitochondrial F_1 in the Protein Data Bank. The crystals were grown in the presence of various inhibitors, nucleotides and transition-state analogues. Although different in detail, 18 structures resemble each other in that only two of the three catalytic sites are occupied by adenine nucleotides, with the third being either empty (in 6) or containing P_i (in 12). In these 18 structures the central stalk has the same orientation, with the bulge of subunit γ facing the empty (or P_i-containing) site. Only in one crystal structure[19] is this site occupied, revealing a half-closed conformation with subunit γ being twisted by 20°.

Stepped rotation

The rotation of subunit γ, the central shaft, relative to the pseudo-hexagon of $(\alpha\beta)_3$ has been time-resolved by single-molecule techniques. ATP hydrolysis drives the stepped rotation of subunit γ with a period of 120° under saturating concentration of MgATP[20]. Under limiting concentration, in the enzymes of thermophilic bacteria and *E. coli*, there are two substeps: the 'ATP-waiting dwell' is followed by a rotation of 80°, and the 'catalytic dwell' by a rotation of 40° (refs 20–22). This assignment is based on the dependence of the respective dwell times on the concentration of ATP and its analogues[23,24], and on experiments in which the nucleotide occupancy[25], conformational changes of one domain on subunit β[26], and the rotation of the central shaft were monitored simultaneously. At least three reactions may be associated with the catalytic dwell: first, the cleavage of bound MgATP into bound MgADP and P_i; second, the release of MgADP; and third, the release of P_i. This final step has been proposed as the rate-determining step of the catalytic dwell, as it supposedly drives the 40° turn[25]. The 80° substep is accompanied by ATP binding[21].

Stepped rotation and crystal structure

The orientations of the central shaft during the transient dwells of the active enzyme are recorded with randomly oriented single molecules. There is no a priori relation between the rotation observed in laboratory coordinates and the molecular coordinates of the crystal structure. This limitation has recently been overcome by recording the stepped rotation of a single molecule and then locking the rotor (subunit γ) by disulphide bridges to the stator (β) in the same orientation as in the above 18 crystal structures[22]. The locked (crystal) orientation of the rotor matches that during the catalytic dwell of the active enzyme, whereas the orientation during the ATP-waiting state is, surprisingly, turned forward by 40° (ref. 22). The respective orientations are illustrated by purple arrows in the reaction scheme in Fig. 2. These results corroborate those inferred from less direct approaches[26–28]. Contrary to the first impression, the crystal structures with two bound adenine nucleotides[10,16] do not seem to represent the ATP-waiting dwell, but rather the catalytic one.

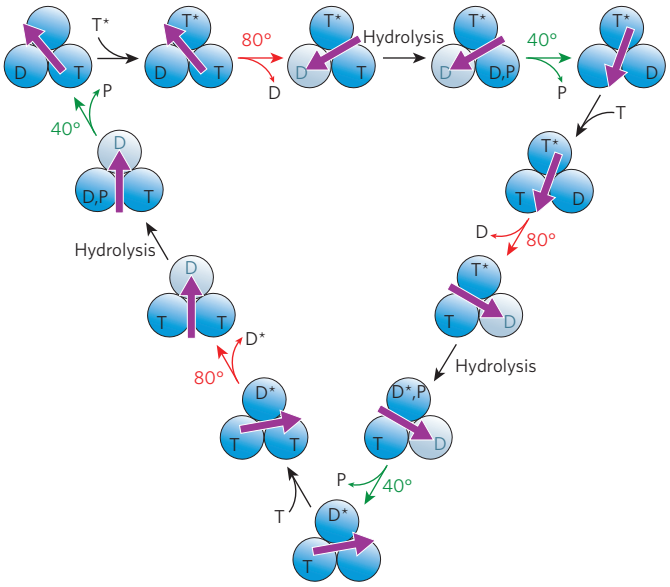


Figure 2 | Tentative reaction scheme of ATP hydrolysis by the chemical rotary motor, F_1. The reaction scheme is arranged as a triangle, each side of which represents 120° of anticlockwise rotation of subunit γ. The purple arrows show the orientation of the bulge of subunit γ relative to the three catalytic sites observed in active *E. coli* F_OF_1[22]. The red and green arrows indicate rotational steps by 80° and 40°, respectively. Starting in the top left-hand corner going clockwise, the one open and empty nucleotide-binding site awaits the binding of an ATP (T), which is called an ATP-waiting dwell. After binding to the receiving site, one molecule of ATP (T*) causes the central shaft to rotate by 80°, accompanied by the release of unlabelled ADP (D) from the second site[25]. Its nucleotide moeity (denoted as T* then D*) remains bound to the enzyme unless the central shaft reaches a position 240° from the start. At least two reactions are associated with the catalytic dwell[24]: cleavage of ATP into ADP and inorganic phosphate (P) in the third site, and the release of P. The release of P is directly coupled to the 40° substep. The enzyme then adopts the next ATP-waiting dwell (top right). The grey sites are either empty (at low concentrations of Mg nucleotides[25]; not indicated here) or filled with loosely bound MgADP (at high concentrations of Mg nucleotides[18]).

Nucleotide occupancy of the catalytic sites

A high turnover rate of the F_1-ATPase involves two, if not all three, of the catalytic sites, which bind MgATP with negative cooperativity. The dissociation constants of *E. coli* F_1 are 1 nM for the first site, 1 μM for the second and about 30 μM for the third[29]. It is still not clear whether two[30] or three[31] sites are occupied by nucleotide during the catalytic dwell. This has resulted in several proposals for the reaction scheme of ATP hydrolysis by F_1[19,25,26,29,30,32–34]. Two seemingly conflicting data sets have emerged, both based on convincing experiments. In one, three-site occupancy for most of the time, with an ATP:ADP ratio of 1:2, has been inferred from experiments with solubilized *E. coli* F_1 under a saturating concentration of MgATP. The occupancy was probed by fluorescent residues facing the catalytic sites[18,34,35]. In the other, two-site occupancy was inferred from single-molecule experiments with F_1 from thermophilic bacteria under a limiting concentration of MgATP, using a fluorescent and hydrolysable ATP derivative as a probe. The label resides on active F_1 only during rotation of γ through 240° (ref. 21), implying that ADP is released during or shortly after ATP binding[25]. Accordingly, two sites are occupied by nucleotide during the catalytic dwell. We propose that the difference between these two data sets is caused by different experimental conditions: saturating MgATP concentration in the first, and non-saturating concentration in the second.

A tentative reaction scheme

The reaction scheme for ATP hydrolysis shown in Fig. 2 accounts for both of the above conflicting data sets if the grey nucleotide-binding site is mostly either occupied or unoccupied, depending on the nucleotide concentration. It implies, however, that the rotary reaction pattern of ATP hydrolysis is largely independent of whether these sites are occupied. The binding of ATP (T in Fig. 2) to the first and empty catalytic site triggers the 80° turn (red arrow) of the central shaft (purple arrow), and simultaneously or shortly thereafter, bound ADP (D) is released from the second site. This is followed by the cleavage of bound ATP into bound ADP and phosphate (P) in the third site. The release of phosphate, which is probably the rate-limiting reaction of the catalytic dwell, triggers the next turn of 40° (green arrow). Under saturating concentrations of ATP, this step limits the overall rate of turnover, causing the intermediate state just before the 40° turn to be detected most of the time. The second site (grey in Fig. 2) is occupied if the concentration of nucleotides is high[18], and empty if it is low[25]. In the latter case, during the catalytic dwell the bulge of subunit γ may point to the empty nucleotide-binding site, as in crystals that were also grown at low nucleotide concentration. Its orientation during the ATP waiting state, 40° forward of this, lacks a correlate in current crystal structures. The scheme in Fig. 2 is based on *ad hoc* assumptions to resolve the conflict and may yet be falsified experimentally.

During the ATP-waiting dwell, one empty site is ready to bind MgATP under conditions of ATP hydrolysis and MgADP under those of synthesis. By separating ATP binding in this site from ADP release by another, depending on the orientation of the central stalk, its rotation is directed[30] anticlockwise (viewed from the membrane side) for ATP hydrolysis[11] and clockwise for synthesis[36].

Torque and thermodynamic efficiency

When a long actin filament is attached to the central shaft of *E. coli* F_OF_1, the viscous drag on the filament slows the turnover rate of the enzyme by several orders of magnitude. The magnitude of the torque produced (M) can then be determined from the elastic deformation of the filament. The observed torque, produced by the stepping motor F_1 and monitored by a probe attached to F_O, varies little as the rotation angle changes[37]. The mean torque $M \approx 56\,\mathrm{pN\,nm}$[37]. Because the enzyme is slowed by viscous drag and operates close to equilibrium, the molar free energy of ATP hydrolysis (ΔG_{ATP}) is expected to equal the magnitude of the molar mechanical work performed on the filament: $M(2\pi/3)N_A = -\Delta G_{ATP}$, where $2\pi/3$ represents the 120° turn of the rotor per ATP molecule hydrolysed, and N_A is Avogadro's number. The above value of the mean torque matches this expectation under given chemical conditions ($\Delta G_{ATP} = -70\,\mathrm{kJ\,mol^{-1}}$)[37]. The match implies 100% efficiency for the conversion of the Gibbs free energy of ATP hydrolysis into mechanical work performed on the elastically strained filament. This is not surprising given the approximate thermodynamic equilibrium of the enzyme (long)-filament construct. It is more informative to say that there is no slip between ATP hydrolysis in F_1 and rotation in F_O under the given conditions. Rotary slip in F_OF_1 in chloroplasts and bacteria has been detected, but only under single-site occupancy, that is, at nucleotide concentrations significantly below 100 nM[38,39]. The momentary torque can be larger (for example, during a particular power stroke) or smaller (during a kinetic dwell) than its equilibrium average. This may account for the still puzzling independence of the torque from the ATP concentration in the nanomolar to millimolar range[40] (see ref. 2 for a review). It is worth mentioning that the other technique for determining the torque from the rate of rotation[11] underestimates its magnitude because it neglects viscous flow coupling between the filament and the enzyme-supporting surface.

Determinants of torque generation by F_1

Many atomic contacts on subunit γ in the $(\alpha\beta)_3$ hexagon have been considered to be involved in torque generation. It is surprising that the axle of subunit γ, which fits into the hydrophobic bearing of $(\alpha\beta)_3$ (Fig. 1b), can be truncated by the deletion of up to 12 amino-acid residues at its C-terminal end in *E. coli* F_1 without a loss of torque[41]. The truncation of 43 residues at the C terminus of subunit γ, plus another 22 at the N terminus in the F_1 of thermophilic bacteria[42], is feasible while retaining some torque production. These treatments restrict torque generation to a narrow ring just above the point where subunit γ protrudes from $(\alpha\beta)_3$ towards F_O (Fig. 1b). This is the contact region between the

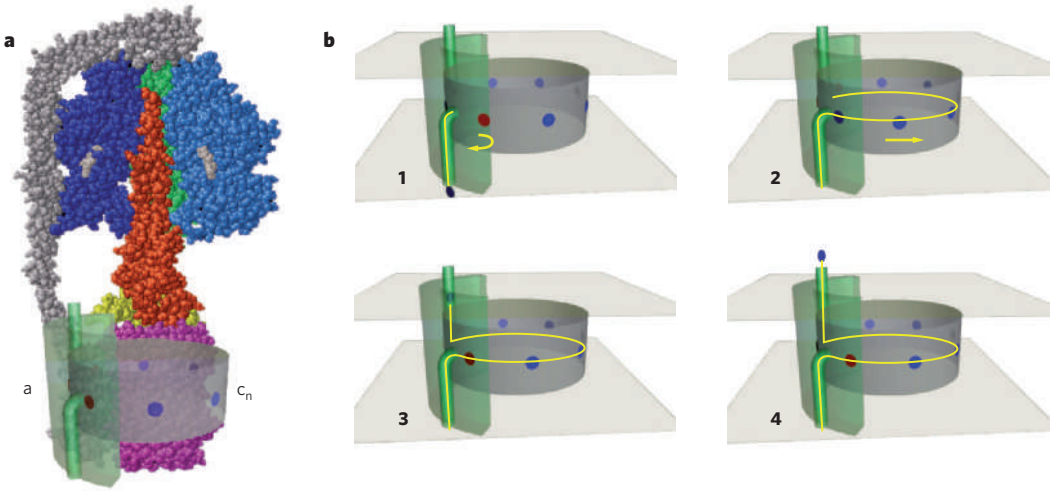


Figure 3 | Model structure illustrating the function of the rotary electromotor, F_O. **a**, An *E. coli* F_OF_1 model (as in Fig. 1) with a schematic representation of the subunit ac_n assembly (see **b**) overlaid. **b**, Four still images from an animation (Supplementary Movie 2) showing torque generation by Brownian rotary motion and directed ion flow. The images show c_n (barrel with spots showing protonated sites in blue and non-protonated and negatively charged sites in red) facing subunit a (green, on the left) with its two laterally offset proton access and exit channels. The path of the proton is indicated by a yellow line. In step 1, c_n rotates back and forth in the plane of the membrane (indicated by the yellow arrow), as the charged site is not protonated and so avoids contact with the hydrophobic centre of the membrane (indicated by planes above and below). In step 2, after protonation has occurred, unidirectional rotation by one step has taken place in the direction indicated by the yellow arrow. Steps 3 and 4 show the proton in the exit channel of subunit a and a return to the initial state by further rotation of c_n.

lever on subunit β and the bulge on γ where the upward and downward motion of the lever (in response to nucleotide binding) pushes the central shaft around, as has been deduced elsewhere[43]. Modelling of the behaviour of F_1 by molecular dynamics alone (see refs 44 and 45, and references therein) has been limited by the narrow width of the time window (about 10 ns). In an attempt to circumvent this complication, and also that of explicitly including the bound substrates and products in the simulation, two different conformations of the $(\alpha\beta)_3$ hexagon have been used as a matrix to rationalize the rotation of the central shaft in F_1 (ref. 46).

Although the full length of subunit γ (Fig. 1) is not needed for torque generation, the rotation of the C-terminal end of γ in the hydrophobic bearing of $(\alpha\beta)_3$ is normally involved in the activity[14,47,48]. As a curiosity, a covalent crosslink between the rotor and the stator in this domain does not impair rotation[49]. A swivel joint created after breaking some hydrogen bonds in the C-terminal α-helical tip of subunit γ seems to be a viable alternative to the normally used axle and bearing.

F_O as a rotary electrochemical motor and generator

Several types of rotary motor are found in nature, but only four are ion-driven: the F_O portion of F_OF_1-ATPase, the A_O portion of A_OA_1-ATPase, the V_O portion of V_OV_1-ATPase, and the flagellar motor. The first three have a similar construction (see refs 50 and 51 for reviews).

A central rotor ring of 10–15 identical copies of the small (8 kDa), hairpin-shaped, hydrophobic subunit c, embedded in a coupling membrane, is a key element of F_O (Fig. 3). The only highly resolved crystal structure of the c-ring has been obtained for a sodium-translocating bacterium, *Ilyobacter tartaricus*[52]. It shows a concave barrel made from 11 copies of subunit c, each with an ion-binding site in the centre of the membrane. This site includes an ion-binding acidic residue. Together with its surrounding pocket, it confers ion selectivity on F_O[15]. During ATP hydrolysis by *E. coli* F_OF_1, the c-ring rotates along with the γε complex, relative to the rest of the enzyme[53,54]. The c-ring is the rotor of F_O, and subunits a and b_2 form the stator (a and b_2 are located at the periphery of the c-ring, and all three subunits are necessary for ion

Box 1 | Terminology

Protonmotive force
The driving force of proton flow across a membrane, this denotes the transmembrane free-energy difference of the proton. It is the sum of the electric (voltage, $\Delta\phi$) and the chemical (entropic) difference ($(-2.3RT/F)\Delta$pH), where R denotes the gas constant, F the Faraday constant and T the temperature. At room temperature and written in millivolts, protonmotive force $\approx \Delta\phi - 59\Delta$pH. In chloroplasts and mitochondria, it can reach 200 mV. The term 'proton gradient', a commonly used synonym for protonmotive force, is a misnomer.

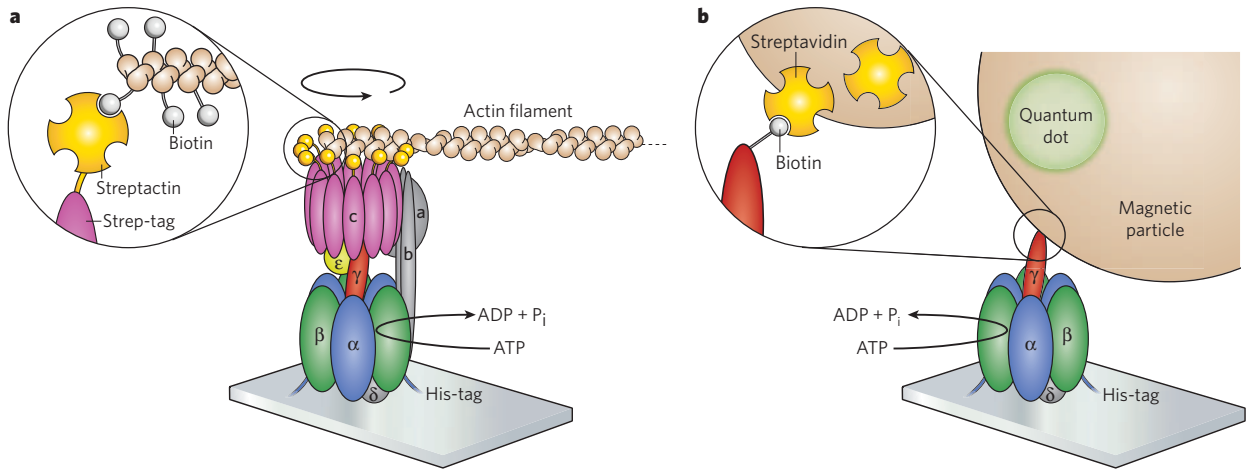
Efficiency
In the context of the ATP synthase, this can be defined in two ways. The thermodynamic efficiency (η_{th}) describes the ratio between the molar free-energy difference (ΔG) of a driving device (ΔG_d) over that of a receiving device (ΔG_r); thus $\eta_{th} = |\Delta G_r/\Delta G_d|$. If both are in equilibrium — that is, if there is no net turnover — the thermodynamic efficiency of the energy transfer between them is maximal ($\eta_{th} = 1$); if they are far from equilibrium, the efficiency may approach zero. The kinetic efficiency (η_{kin}) relates the actual reaction flux (J_{act}) to the maximum possible one (J_{max}), thus $\eta_{kin} = J_{act}/J_{max}$. This depends on the design of a given engine.

Elastic power transmission
Is an elastic element between a motor and another mechanical device that can transiently store and transmit elastic deformation energy. The stepped power strokes of the motor are then smoothed, such that the other device is exposed to almost constant force or torque. The elastic element is characterized by its spring constant (κ).

Fluctuation analysis
Every molecular system is subject to thermal collisions with its neighbours. Every elastic system, agitated by those impacts, fluctuates around its equilibrium position and, on average, stores a mean elastic energy of magnitude k_BT, where k_B is the Boltzmann constant and T is the temperature in kelvins. The fluctuation profile of elastic devices is Gaussian. Its variance, σ, yields the elastic spring constant κ according to $\kappa = k_BT/\sigma$ (measured in newton metres)[74].

Torque
Torque (M) = force × lever-arm-length (in N m or, for nano-machines, in pN nm). It is the equivalent in rotary motion of force in linear motion. When applied to a spinning top it accelerates rotation, and to a flexible rod with one side fixed it causes torsion.

Box 2 | Measurements of rotation, torque and efficiency



Rotation of subunit γ, the central shaft in the chemical nanomotor F_1, relative to the $(\alpha\beta)_3$ barrel is compatible with biochemical crosslinking data[76]. It received initial support from time-resolved polarized absorption photometry[47], but unidirectional rotation could not at first be demonstrated unequivocally. It was spectacularly proven in single-molecule experiments using the F_1 from the thermophile *Bacillus PS3* (ref. 11). The image illustrates the principle, here applied to *E. coli* F_OF_1 (**a**) and F_1 (**b**). The engineered enzyme is fixed by histidine (His) tags to a nickel-coated solid support. A fluorophore-labelled actin filament (in **a**, attached to the c ring on F_OF_1) or a quantum-dot-labelled hyper-paramagnetic bead (in **b**, attached to the C-terminal end of subunit γ in F_1) serve to monitor the ATP-driven rotation in a fluorescence microscope. Those beads can also be used to drive the activity mechanically by a rotating magnetic field, and one team eventually managed to drive ATP synthesis by F_1 magnetically in *Bacillus PS3* (ref. *77*). Whereas the yield of magnetically induced rotation can be almost 100% of all labelled molecules in a given viewing field over several hours[74], the yield is usually less than 1% sustained for just a few minutes when driven by the hydrolysis of ATP. Here, rotating enzyme molecules are a highly selected subset. Because they are attached to the surface, their kinetic parameters may differ from those in solution. The possibility of distortions of this particularly mobile enzyme, both by the immobilization of single molecules and in crystals, has to be considered.

Nano-sized gold beads (diameter 30 nm)[23] or small filaments (typical length 0.4 μm)[74] have been used to monitor the kinetic details of stepped rotation. Because the viscous drag on a rotating filament scales with the third power of its length[78], long filaments (say, 3 μm long) slow the enzyme turnover more than 1,000-fold, such that the enzyme–filament system operates close to thermodynamic equilibrium. Under these conditions, the rotating filament is elastically bent by the enzyme and viscous torque, and can be used as a spring balance to determine the torque produced by F_1 and to assess the enzyme's thermodynamic efficiency[37].

The above techniques are limited to immobilized single molecules, but Förster resonance energy transfer between two dye molecules linked to rotor and stator has allowed the rotary progression in F_OF_1 to be monitored after reconstitution in proteoliposomes. Stepped inter-subunit rotation has been observed both in one direction during ATP hydrolysis and in the opposite direction during proton-driven ATP synthesis[36] (see also refs 79 and 80). (Panel **a** is modified with permission from ref. 53; **b** is modified with permission from ref. 74.)

transport[55]). Subunit a has five transmembrane helices and provides a positively charged residue (Arg 210 in *E. coli*) and probably the ion access channels at the interface with the c-ring[56–58]. Subunit b is elongated and largely α-helical. Anchored to the membrane by its N-terminal hydrophobic domain (see ref. 59 for a review), it extends up and binds with high affinity to $(\alpha\beta)_3$, both directly and through δ[60–62].

A hypothesis for torque generation

How ion flow through F_O generates the torque needed to power F_1 has been debated since 1993. The current concept[12,13,63] is shown in Fig. 3b in the form of four still images from a video animation (Supplementary Movie 2). A rigid ring of 10 c subunits (in *E. coli*) is thought to carry out Brownian rotational fluctuations relative to the a subunit of the stator of F_O (green in Fig. 3b). The range of symmetrical fluctuations is restricted by two electrostatic constraints: the acidic residue must be negatively charged (red dots in Fig. 3b) when facing the electropositive stator, and neutral (protonated) when facing the hydrophobic lipid (blue dots in Fig. 3b). Two access channels for ions connect the ion-binding site with either side of the membrane. When a proton (blue dots in Fig. 3b) enters from the lower access channel and binds to the acid residue, it relieves the electrostatic constraint, so the fluctuating ring can move one step anticlockwise. Another proton then dissociates from its binding site on the c-ring and exits through the upper channel (Fig. 3b, step 4). The non-collinear arrangement of these channels confers chirality. The direction of rotary motion is determined as follows: the c-ring rotates anticlockwise if the activity is higher on the lower side of the membrane than on the upper one, and clockwise if the activity difference is reversed. In this way, ion flow is used to generate torque. This simple concept has survived the arrival of a crystal structure of the c-ring[64], whereas postulated modifications concerning the access-channel configuration[65] and more complicated dynamics of the c-ring[66,67] are not supported (see ref. 51 and references therein).

Unitary proton conductance

The unitary conductance of *E. coli* F_O has been determined experimentally[68]. The time-averaged conductance is 10 fS, which is equivalent to the translocation of 6,500 protons per second at an electric driving force of 100 mV. The conductance of bare F_O is at least tenfold larger than that of integral F_OF_1. The current–voltage relation is linear (ohmic). In the absence of F_1, F_O is not voltage-gated. The proton specificity is extremely high (the rejection ratio for H^+ over K^+ and Na^+ is greater than 10^7), and conductance depends only mildly on pH (a factor of only two between pH 6.5 and pH 10). The simple rotary transport model shown in Fig. 3 is compatible with these data, as has been revealed in a simulation by statistical mechanics that requires only three free parameters[68], in particular, different pK-values of 6.5 and 10 of the two access channels.

F_OF_1 is a coupled double motor and generator

F_O and F_1 have different gear ratios in different organisms. Atomic force microscopy and X-ray crystal-structure analysis have revealed that the ring structure in F_O may consist of 10, 11, 13, 14 or 15 copies of subunit c, depending on the organism (see ref. 64 and references therein). The

structural gear ratio (the number of copies of subunit c in the ring over three catalytic sites) therefore differs between species, but not within a species as previous reports of a nutrition-dependent variation of this number in two organisms have not been verified[69]. The presence of 10, 11, 13, 14 or 15 copies of subunit c implies proton:ATP ratios of 3.3, 3.7, 4.3, 4.7 and 5.0. This has not been found. For chloroplast F_OF_1, for instance, several laboratories have consistently found a proton:ATP ratio of 4.0 (see ref. 70 and references therein), although a ratio of 4.7 would be expected as there are 14 monomers in the c-ring[71]. A larger ratio than expected could imply imperfect coupling, but a smaller one, as in this example, is difficult to rationalize.

F_OF_1 can operate in either direction, depending on which driving force dominates (unless it is downregulated for ATP hydrolysis, as in chloroplasts). Its rotary electromotor has ten steps per round in *E. coli* and drives the chemical generator to produce ATP in three or six steps, or *vice versa*. Are their respective power strokes fine-tuned to one another? Alternatively, it has been claimed that an elastic power transmission between the two motors smoothes their cooperation without the need to fine-tune their detailed molecular events[13,72,73]. The first direct evidence for elastic coupling was the almost constant torque delivered to F_O in *E. coli*, even though the driving motor (F_1) is a stepping motor[37]. Simulation of the coupled motors by the Fokker–Planck equation of statistical mechanics has reproduced the smoothing of the torque output in the presence of an elastic element between F_O and F_1, with a torsional spring constant (see Box 1) of about 60 pN nm. It has also shown that an elastic power transmission is indispensable for a high rate of coupled turnover under load. It increases the rate by several orders of magnitude over that of a rigidly coupled double motor[37].

Which domains might be responsible for the elastic power transmission? This question has recently been addressed by 'fluctuation analysis' (see Box 1) for single molecules of *E. coli* F_OF_1. The elastic compliance of certain domains was restricted by the engineering of disulphide bridges between the rotor and the stator, and the elasticity of the unrestricted domains was determined from the width of the thermal rotary fluctuations[74]. The main body of the enzyme, $(\alpha\beta)_3$, and the eccentric stator are very stiff, as is the portion of subunit γ that is inserted into $(\alpha\beta)_3$ (see Box 2). Only the portion between F_O and F_1, involving subunits γ, ε and c_{10}, is more flexible, by one order of magnitude ($\kappa \approx 70$ pN nm). If this domain is exposed to the mean torque of 56 pN nm, developed in F_1 by ATP hydrolysis and countered by F_O, this domain will be twisted by about 46°, storing about 14 kJ mol^{-1} of elastic energy. The flexible lever on subunit β adds the same amount of elastic flexibility, reducing the effective spring constant of the active enzyme to about 35 pN nm. The elastic power transmission both increases the 'kinetic efficiency' of the coupled motors (see Box 1) and allows the double motor to function with different gears in different organisms[74].

Outlook

ATP synthase (F_OF_1) is a molecular machine that combines the electrical, mechanical and chemical aspects of enzyme function. These are neatly separated, readily attributed to its different subunits, and reasonably well understood thanks to a wealth of structural and kinetic data. However, understanding the enzyme fully at a molecular level will require considerable efforts, both experimental and theoretical. There are five outstanding issues. First, high-resolution structures of complete F_OF_1 are required. Its elastic flexibility and complex structure mean that F_OF_1 is not an easy target for crystallization, especially in defined states of catalysis. Second, full kinetic and mechanistic characterization of the partial reactions on the way from ADP to ATP are needed. The present data have been obtained using enzymes from different organisms, which complicates matters. Third, more information is needed about the mechanism of proton translocation through F_O. This requires more detailed structural information than is currently available. Fourth, the inconsistency between structurally expected and thermodynamically or kinetically observed proton:ATP ratios must be explained, as this is fundamental to our understanding of ATP synthesis. Finally, new theoretical approaches must be developed to extend the description from the nanosecond time domain of molecular mechanics to the level of milliseconds. This will provide fresh mechanistic insights and pave the way for new experimental approaches. Only when we have solved these problems will we come close to a full understanding of this remarkable piece of cellular machinery. ■

1. Mitchell, P. Coupling of photophosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. *Nature* **191,** 144–148 (1961).
In this seminal paper, Peter Mitchell introduces the 'chemiosmotic hypothesis', his concept of ion-coupled ATP synthesis.
2. Junge, W. in *Photosynthesis* (ed. Renger, G.) 447–467 (Royal Society of Chemistry, 2008).
3. Krulwich, T. A. *et al.* Energetics of alkaliphilic *Bacillus* species: physiology and molecules. *Adv. Microb. Physiol.* **40,** 401–438 (1998).
4. Cherepanov, D. A. *et al.* Low dielectric permittivity of water at the membrane interface: effect on the energy coupling mechanism in biological membranes. *Biophys. J.* **85,** 1307–1316 (2003).
5. Penefsky, H. S. *et al.* Partial resolution of the enzymes catalyzing oxidative phosphorylation. II. Participation of a soluble adenosine triphosphatase in oxidative phosphorylation. *J. Biol. Chem.* **235,** 3330–3336 (1960).
6. Boyer, P. D., Cross, R. L. & Momsen, W. A new concept for energy coupling in oxidative phosphorylation based on a molecular explanation of the oxygen exchange reactions. *Proc. Natl Acad. Sci. USA* **70,** 2837–2839 (1973).
This is the first article in a series in which Paul Boyer emphasizes the cooperation of at least two equivalent reaction sites in the F_1-ATPase, paving the way to the present view of a rotary 'binding change mechanism'.
7. Boyer, P. D. in *Membrane Bioenergetics* (eds Lee, C. P., Schatz, G. & Ernster, L.) 476–479 (Addison-Wesley, 1979).
8. Boyer, P. D. The ATP synthase – a splendid molecular machine. *Annu. Rev. Biochem.* **66,** 717–749 (1997).
9. Boyer, P. D. & Kohlbrenner, W. E. in *Energy Coupling in Photosynthesis* (eds Selman, B. R. & Selman-Reimer, S.) 231–241 (Elsevier, 1981).
10. Abrahams, J. P. *et al.* The structure of F_1-ATPase from bovine heart mitochondria determined at 2.8 Å resolution. *Nature* **370,** 621–628 (1994).
John Walker and colleagues present the first of several asymmetric crystal structures of the F_1-ATPase and outline the structural features of rotary catalysis.
11. Noji, H. *et al.* Direct observation of the rotation of F_1-ATPase. *Nature* **386,** 299–302 (1997).
This is the pioneering article in an impressive and technically highly refined series in which the Japanese teams use microvideography to record the rotation in F_1 in real time.
12. Vik, S. B. & Antonio, B. J. A mechanism of proton translocation by F_1F_O ATP synthases suggested by double mutants of the a subunit. *J. Biol. Chem.* **269,** 30364–30369 (1994).
13. Junge, W., Lill, H. & Engelbrecht, S. ATP synthase: an electrochemical transducer with rotatory mechanics. *Trends Biochem. Sci.* **22,** 420–423 (1997).
This article outlines how rotary ion conduction by F_O generates torque as a result of Brownian inter-subunit rotation and electrostatic constraints.
14. Sabbert, D., Engelbrecht, S. & Junge, W. Functional and idling rotatory motion within F-ATPase. *Proc. Natl Acad. Sci. USA* **94,** 4401–4405 (1997).
15. von Ballmoos, C., Cook, G. M. & Dimroth, P. Unique rotary ATP synthase and its biological diversity. *Annu. Rev. Biophys.* **37,** 43–64 (2008).
16. Bowler, M. W. *et al.* Ground state structure of F_1-ATPase from bovine heart mitochondria at 1.9 A resolution. *J. Biol. Chem.* **282,** 14238–14242 (2007).
17. Weber, J. *et al.* α-Aspartate 261 is a key residue in noncatalytic sites of *Escherichia coli* F_1-ATPase. *J. Biol. Chem.* **270,** 21045-21049 (1995).
18. Weber, J., Bowman, C. & Senior, A. E. Specific tryptophan substitution in catalytic sites of *Escherichia coli* F_1-ATPase allows differentiation between bound substrate ATP and product ADP in steady-state catalysis. *J. Biol. Chem.* **271,** 18711–18718 (1996).
Alan Senior and colleagues introduce a technique to monitor the nucleotide occupancy of the catalytic sites on F_1 by the fluorescence of engineered tryptophan residues.
19. Menz, R. I., Walker, J. E. & Leslie, A. G. Structure of bovine mitochondrial F_1-ATPase with nucleotide bound to all three catalytic sites: implications for the mechanism of rotary catalysis. *Cell* **106,** 331–341 (2001).
20. Yasuda, R. *et al.* Resolution of distinct rotational substeps by submillisecond kinetic analysis of F_1-ATPase. *Nature* **410,** 898–904 (2001).
21. Nishizaka, T. *et al.* Chemomechanical coupling in F_1-ATPase revealed by simultaneous observation of nucleotide kinetics and rotation. *Nature Struct. Mol. Biol.* **11,** 142–148 (2004).
22. Sielaff, H. *et al.* Functional halt positions of rotary F_OF_1-ATPase correlated with crystal structures. *Biophys. J.* **95,** 4979–4987 (2008).
23. Yasuda, R. *et al.* F_1-ATPase is a highly efficient molecular motor that rotates with discrete 120° steps. *Cell* **93,** 1117–1124 (1998).
24. Shimabukuro, K. *et al.* Catalysis and rotation of F_1 motor: cleavage of ATP at the catalytic site occurs in 1 ms before 40° substep rotation. *Proc. Natl Acad. Sci. USA* **100,** 14731–14736 (2003).
25. Adachi, K. *et al.* Coupling of rotation and catalysis in F_1-ATPase revealed by single-molecule imaging and manipulation. *Cell* **130,** 309–321 (2007).
26. Masaike, T. *et al.* Cooperative three-step motions in catalytic subunits of F_1-ATPase correlate with 80° and 40° substep rotations. *Nature Struct. Mol. Biol.* **15,** 1326–1333 (2008).
27. Hirono-Hara, Y. *et al.* Pause and rotation of F_1-ATPase during catalysis. *Proc. Natl Acad. Sci. USA* **98,** 13649–13654 (2001).
28. Yasuda, R. *et al.* The ATP-waiting conformation of rotating F_1-ATPase revealed by single-pair fluorescence resonance energy transfer. *Proc. Natl Acad. Sci. USA* **100,** 9314–9318 (2003).
29. Senior, A. E., Nadanaciva, S. & Weber, J. The molecular mechanism of ATP synthesis by F_1F_O-ATP synthase. *Biochim. Biophys. Acta* **1553,** 188–211 (2002).
30. Boyer, P. D. Catalytic site occupancy during ATP synthase catalysis. *FEBS Lett.* **512,** 29–32 (2002).
31. Weber, J. & Senior, A. E. ATP synthase: what we know about ATP hydrolysis and what we

do not know about ATP synthesis. *Biochim. Biophys. Acta* **1458,** 300–309 (2000).
32. Ariga, T., Muneyuki, E. & Yoshida, M. F_1-ATPase rotates by an asymmetric, sequential mechanism using all three catalytic subunits. *Nature Struct. Mol. Biol.* **14,** 841–846 (2007).
33. Nakamoto, R. K., Baylis Scanlon, J. A. & Al-Shawi, M. K. The rotary mechanism of the ATP synthase. *Arch. Biochem. Biophys.* **476,** 43–50 (2008).
34. Mao, H. Z. & Weber, J. Identification of the β_{TP} site in the X-ray structure of F_1-ATPase as the high-affinity catalytic site. *Proc. Natl Acad. Sci. USA* **104,** 18478–18483 (2007).
35. Weber, J. *et al.* Specific placement of tryptophan in the catalytic sites of *Escherichia coli* F_1-ATPase provides a direct probe of nucleotide binding: maximal ATP hydrolysis occurs with three sites occupied. *J. Biol. Chem.* **268,** 20126–20133 (1993).
36. Diez, M. *et al.* Proton-powered subunit rotation in single membrane-bound F_OF_1-ATP synthase. *Nature Struct. Mol. Biol.* **11,** 135–141 (2004).
37. Pänke, O. *et al.* Viscoelastic dynamics of actin filaments coupled to rotary F-ATPase: torque profile of the enzyme. *Biophys. J.* **81,** 1220–1233 (2001).
38. Feniouk, B. A., Mulkidjanian, A. Y. & Junge, W. Proton slip in the ATP synthase of *Rhodobacter capsulatus*: induction, proton conduction, and nucleotide dependence. *Biophys. Biochim. Acta* **1706,** 184–194 (2005).
39. Groth, G. & Junge, W. Proton slip of chloroplast ATPase: its nucleotide dependence, energetic threshold and relation to an alternating site mechanism of catalysis. *Biochemistry* **32,** 8103–8111 (1993).
40. Sakaki, N. *et al.* One rotary mechanism for F_1-ATPase over ATP concentrations from millimolar down to nanomolar. *Biophys. J.* **88,** 2047–2056 (2005).
41. Müller, M. *et al.* F_1-ATPase: the C-terminal end of subunit γ is not required for ATP hydrolysis-driven rotation. *J. Biol. Chem.* **277,** 23308–22313 (2002).
42. Furuike, S. *et al.* Axle-less F_1-ATPase rotates in the correct direction. *Science* **319,** 955–958 (2008).
43. Sun, S. *et al.* Elastic energy storage in β-sheets with application to F_1-ATPase. *Eur. Biophys. J.* **32,** 676–683 (2003).
44. Dittrich, M. & Schulten, K. Zooming in on ATP hydrolysis in F_1. *J. Bioenerget. Biomembr.* **37,** 441–444 (2005).
45. Bockmann, R. A. & Grubmuller, H. Conformational dynamics of the F_1-ATPase β-subunit: a molecular dynamics study. *Biophys. J.* **85,** 1482–1491 (2003).
46. Pu, J. & Karplus, M. How subunit coupling produces the γ-subunit rotary motion in F_1-ATPase. *Proc. Natl Acad. Sci. USA* **105,** 1192–1197 (2008).
47. Sabbert, D., Engelbrecht, S. & Junge, W. Intersubunit rotation in active F-ATPase. *Nature* **381,** 623–625 (1996).
48. Müller, M. *et al.* Rotary F_1-ATPase. Is the C-terminus of subunit γ fixed or mobile? *Eur. J. Biochem.* **271,** 3914–3922 (2004).
49. Gumbiowski, K. *et al.* F-ATPase: forced full rotation of the rotor despite covalent cross-link with the stator. *J. Biol. Chem.* **276,** 42287–42292 (2001).
50. Feniouk, B. & Junge, W. in *The Purple Bacteria* (eds Hunter, C. L. *et al.*) 475-493 (Springer, 2008).
51. Junge, W. & Nelson, N. Nature's rotary electromotors. *Science* **308,** 642–644 (2005).
52. Meier, T. *et al.* Structure of the rotor ring of F-type Na^+-ATPase from *Ilyobacter tartaricus*. *Science* **308,** 659–662 (2005).
53. Pänke, O. *et al.* F-ATPase: specific observation of the rotating c subunit oligomer of EF_OEF_1. *FEBS Lett.* **472,** 34–38 (2000).
54. Sambongi, Y. *et al.* Mechanical rotation of the c subunit oligomer in ATP synthase (F_OF_1): direct observation. *Science* **286,** 1722–1724 (1999).
55. Schneider, E. & Altendorf, K. All three subunits are required for the reconstitution of an active proton channel (F_O) of *Escherichia coli* ATP synthase (F_1F_O). *EMBO J.* **4,** 515–518 (1985).
56. Angevine, C. M., Herold, K. A. & Fillingame, R. H. Aqueous access pathways in subunit a of rotary ATP synthase extend to both sides of the membrane. *Proc. Natl Acad. Sci. USA* **100,** 13179–13183 (2003).
57. Zhang, D. & Vik, S. B. Helix packing in subunit a of the *Escherichia coli* ATP synthase as determined by chemical labeling and proteolysis of the cysteine-substituted protein. *Biochemistry* **42,** 331–337 (2003).
58. Langemeyer, L. & Engelbrecht, S. Essential arginine in subunit a and aspartate in subunit c of F_OF_1 ATP synthase. Effect of repositioning within helix 4 of subunit a and helix 2 of subunit c. *Biochim. Biophys. Acta* **1767,** 998–1005 (2007).
59. Dunn, S. D. *et al.* The b subunit of *Escherichia coli* ATP synthase. *J. Bioenerget. Biomembr.* **32,** 347–355 (2000).
60. Diez, M. *et al.* Binding of the b-subunit in the ATP synthase from *Escherichia coli*. *Biochemistry* **43,** 1054–1064 (2004).
61. Weber, J. *et al.* Quantitative determination of direct binding of b subunit to F_1 in *Escherichia coli* F_1F_O-ATP synthase. *J. Biol. Chem.* **279,** 11253–11258 (2004).
62. Häsler, K., Pänke, O. & Junge, W. On the stator of rotary ATP synthase: the binding strength of subunit δ to $(ab)_3$ as determined by fluorescence correlation spectroscopy. *Biochemistry* **38,** 13759–13765 (1999).
63. Junge, W. On Vik's letter concerning comments on a statement in W. Junge (2004). *Photosynth. Res.* **87,** 233 (2006).
64. Meier, T. *et al.* A tridecameric c ring of the adenosine triphosphate (ATP) synthase from the thermoalkaliphilic *Bacillus* sp. strain TA2.A1 facilitates ATP synthesis at low electrochemical proton potential. *Mol. Microbiol.* **65,** 1181–1192 (2007).
65. Dimroth, P. *et al.* Energy transduction in the sodium F-ATPase of *Propionigenium modestum*. *Proc. Natl Acad. Sci. USA* **96,** 4924–4928 (1999).
66. Fillingame, R. H., Angevine, C. M. & Dmitriev, O. Y. Mechanics of coupling proton movements to c-ring rotation in ATP synthase. *FEBS Lett.* **555,** 29–34 (2003).
67. Fillingame, R. H., Angevine, C. M. & Dmitriev, O. Y. Coupling proton movements to c-ring rotation in F_1F_O ATP synthase: aqueous access channels and helix rotations at the a–c interface. *Biochim. Biophys. Acta* **1555,** 29–36 (2002).
68. Feniouk, B. A. *et al.* The proton driven rotor of ATP synthase: ohmic conductance (10 fS) and absence of voltage gating. *Biophys. J.* **86,** 4094–4109 (2004).
69. Ballhausen, B., Altendorf, K. & Deckers-Hebestreit, G. Constant c_{10} ring stoichiometry in the *Escherichia coli* ATP synthase analyzed by cross-linking. *J. Bacteriol.* **191,** 2400–2404 (2009).
70. Steigmiller, S., Turina, P. & Gräber, P. The thermodynamic H^+/ATP ratios of the H^+-ATP-synthases from chloroplasts and *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105,** 3745–3750 (2008).
71. Seelert, H., Dencher, N. A. & Müller, D. J. Fourteen protomers compose the oligomer III of the proton-rotor in spinach chloroplast ATP synthase. *J. Mol. Biol.* **333,** 337–344 (2003).
72. Cherepanov, D. A., Mulkidjanian, A. & Junge, W. Transient accumulation of elastic energy in proton translocating ATP synthase. *FEBS Lett.* **449,** 1–6 (1999).
73. Pänke, O. & Rumberg, B. Kinetic modeling of rotary CF_OF_1-ATP synthase: storage of elastic energy during energy transduction. *Biochim. Biophys. Acta* **1412,** 118–128 (1999).
74. Sielaff, H. *et al.* Domain compliance and elastic power transmission in rotary F_OF_1-ATPase. *Proc. Natl Acad. Sci. USA* **105,** 17760–17765 (2008).
This is the latest article in a series in which the necessary elastic power transmission between the stepping motors F_O and F_1 is correlated with the elastic compliance of specific domains.
75. RasMol v.2.7.2.1.1 (<http://www.rasmol.org>, 2004).
76. Duncan, T. M. *et al.* Rotation of subunits during catalysis by *Escherichia coli* F_1- ATPase. *Proc. Natl Acad. Sci. USA* **92,** 10964–10968 (1995).
77. Itoh, H. *et al.* Mechanically driven ATP synthesis by F_1-ATPase. *Nature* **427,** 465–468 (2004).
78. Cherepanov, D. A. & Junge, W. Viscoelastic dynamics of actin filaments coupled to rotary F-ATPase: curvature as an indicator of the torque. *Biophys. J.* **81,** 1234–1244 (2001).
79. Böckmann, R. A. & Grubmüller, H. Nanosecond molecular dynamics simulation of primary mechanical energy transfer steps in F_1-ATP synthase. *Nature Struct. Biol.* **9,** 198–202 (2002).
80. Ma, J. *et al.* A dynamical analysis of the rotation mechanism for conformational change in F_1-ATPase. *Structure* **10,** 921–931 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Cherepanov and O. Pänke for their contributions to elastic coupling, G. Hikade and H. Kenneweg for technical assistance, and J. Weber for discussions. We acknowledge financial support from Deutsche Forschungsgemeinschaft, Volkswagenstiftung, European Union and Fonds der Chemie.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to W.J. (junge@uos.de).

How intramembrane proteases bury hydrolytic reactions in the membrane

Elinor Erez[1], Deborah Fass[2] & Eitan Bibi[1]

Intramembrane proteolysis is increasingly seen as a regulatory step in a range of diverse processes, including development, organelle shaping, metabolism, pathogenicity and degenerative disease. Initial scepticism over the existence of intramembrane proteases was soon replaced by intense exploration of their catalytic mechanisms, substrate specificities, regulation and structures. Crystal structures of metal-dependent and serine intramembrane proteases have revealed active sites embedded in the plane of the membrane but accessible by water, a requirement for hydrolytic reactions. Efforts to understand how these membrane-bound proteases carry out their reactions have started to yield results.

Proteases have been studied by molecular biologists for decades. The first diffraction patterns of a globular protein were collected from crystals of the protease pepsin[1], and proteases or their inhibitors constituted five of the first dozen members of the Protein Data Bank[2]. Proteases are targets for numerous drugs[3], teaching tools for introductory enzymology and part of a molecular biologist's tool kit. They are grouped into four main classes: serine, cysteine, aspartyl and metalloproteases (Box 1). Enzymes in some of these classes use water to attack a peptide bond as the first step of substrate cleavage; others use water later in the cleavage mechanism, to resolve a covalent intermediate formed between the enzyme and the substrate.

The study of proteases entered a new phase, literally, with the discovery of intramembrane-cleaving proteases (I-CLiPs)[4], which catalyse peptide bond cleavage in membranes. Such hydrolytic reactions were previously thought to be limited to water-soluble enzymes. Intramembrane proteolysis provides a simple and irreversible strategy for signalling: the cleavage and release of membrane-tethered protein domains. I-CLiP activity can liberate cytosolic domains that enter the nucleus to act as transcription factors[5] or liberate external domains to activate receptors on neighbouring cells[6]. So far, I-CLiPs have been identified that belong to the serine, aspartyl and metalloprotease classes, but the existence of cysteine I-CLiPs remains an open question.

In recent years, great progress has been made in the study of I-CLiPs, leading to a better understanding of the concept of intramembrane proteolysis. This Review combines the available insights into the structures, mechanisms of catalysis and physiological roles of I-CLiPs. Although we now understand how these enzymes span the membrane in such a way as to allow water to enter the buried active site, the mode of interaction with membrane protein substrates remains unclear.

Intramembrane proteases

Intramembrane proteolysis was first described as an essential activity in sterol homeostasis[4]. The cytoplasmic domain of the mammalian sterol regulatory element-binding protein (SREBP) is initially tethered by a transmembrane segment to the membrane of the endoplasmic reticulum (ER). When cholesterol levels in the cell drop, SREBP is trafficked to the Golgi apparatus, where it encounters an intramembrane metalloprotease that cleaves the SREBP cytoplasmic domain from its tether. The liberated domain then travels to the nucleus, where it switches on the genes required for cholesterol and fatty-acid synthesis[5]. Since this breakthrough in establishing the principle of intramembrane proteolysis, more I-CLiPs have been identified, making up three major groups (Fig. 1 and Table 1): S2P zinc metalloproteases; rhomboid serine proteases; and signal peptide peptidase (SPP) and presenilin (PSEN), two related aspartyl proteases. The following sections will introduce these I-CLiP groups and describe representative physiological functions.

Intramembrane metalloproteases

The metal-dependent I-CLiP that cleaves SREBP was named site-2 protease (S2P), reflecting the requirement for preprocessing of the substrate by another enzyme, termed site-1 protease (S1P)[4]. Proteases of the S2P family (Fig. 1a) have been identified on the basis of sequence similarity in eukaryotes, bacteria and archaea[5,7,8]. They contain at least four transmembrane helices and a HEXXHX_nDG metal-binding motif. In addition to signalling that there is insufficient cholesterol and taking part in fatty acid and lipid homeostasis, S2P metalloproteases participate in the responses to protein misfolding in both the mammalian ER lumen and the bacterial periplasm. The presence of unfolded proteins in a stressed ER induces the trafficking of activating transcription factor 6 (ATF6) to the Golgi apparatus, where it is processed by S1P and S2P[9]. The released cytosolic domain of ATF6 induces the expression of factors that alleviate ER stress. Similarly, failure to properly assemble and integrate proteins into the outer membrane of the bacterium *Escherichia coli* is reported by a proteolytic cascade involving an S2P family I-CLiP[10]. Additional forms of stress response, including sporulation and the reaction to cell-envelope insults in the bacterium *Bacillus subtilis*, also involve S2P proteases[8,11]. Phylogenetic analysis and conceptual similarities in the present-day function of these enzymes suggest that intramembrane metalloproteases were used in an ancient mechanism for sensing and reporting extracytoplasmic stress[12].

Intramembrane serine proteases

The intramembrane serine proteases are also widely spread across eukaryotes, Bacteria and Archaea[13,14]. These enzymes contain at least six transmembrane helices and the active-site motif GXSX_nH. Unlike S2P proteases, serine I-CLiPs do not require S1P-mediated precleavage of their substrates, and they often participate in intercellular rather than intracellular signalling. Serine I-CLiPs, termed 'rhomboid' proteases

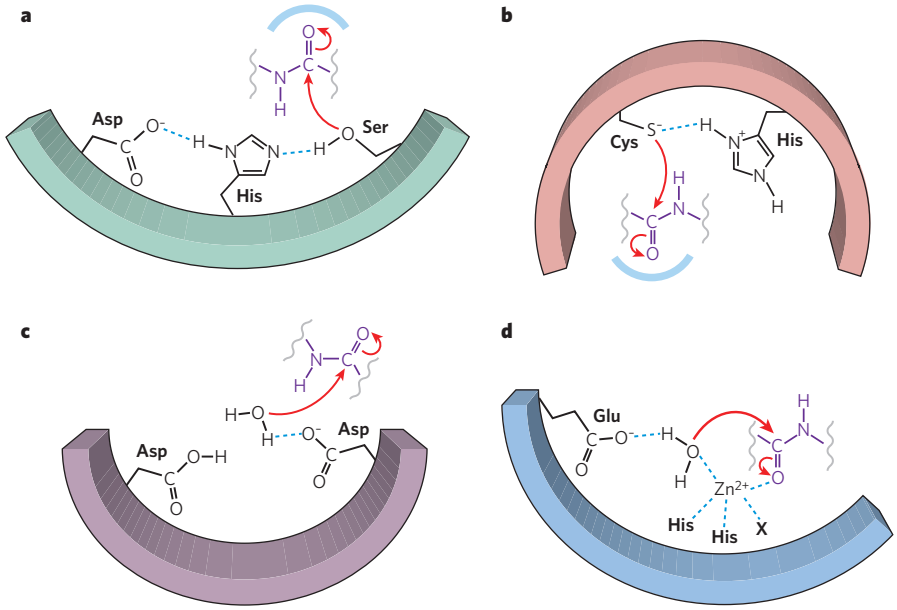
[1]Department of Biological Chemistry, [2]Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel.

Box 1 | Protease mechanisms

Polypeptides can be cleaved either chemically or enzymatically. Enzymes that catalyse the hydrolytic cleavage of peptide bonds are called proteases. Proteases fall into four main mechanistic classes: serine, cysteine, aspartyl and metalloproteases. In the active sites of serine and cysteine proteases, the eponymous residue is usually paired with a proton-withdrawing group to promote nucleophilic attack on the peptide bond. Aspartyl proteases and metalloproteases activate a water molecule to serve as the nucleophile, rather than using a functional group of the enzyme itself. However, the overall process of peptide bond scission is essentially the same for all protease classes.

Soluble serine proteases (EC 3.4.21; **a** in the image) use a catalytic triad located in the active site of the enzyme. The triad is a coordinated structure consisting of three amino acids: histidine, serine and aspartic acid. Each of these three amino acids has an important role in the protease mechanism. The histidine, with the aid of the proton-withdrawing aspartate, deprotonates the serine hydroxyl, enabling nucleophilic attack on the substrate carbonyl carbon. However, intramembrane serine proteases lack the active-site aspartic acid and seem to function with a catalytic dyad instead. The second characteristic feature of serine proteases is an oxyanion hole, which stabilizes the tetrahedral transition state. In the serine protease chymotrypsin, for example, the oxyanion hole is formed using the backbone NH groups of glycine 193 and serine 195.

The mechanism of cysteine proteases (EC 3.4.22; **b** in the image) is similar to that of serine proteases in the use of a strong nucleophile and the formation of a covalent enzyme–substrate complex. However, the nucleophile is the sulphur atom of a cysteine residue, as opposed to the oxygen atom of a serine. In the cysteine protease papain, the backbone NH groups of cysteine 25 and glutamine 19 form the oxyanion hole.



The aspartyl proteases (EC 3.4.23; **c** in the image) contain two aspartic acid residues that act in a general acid–base mechanism. A water molecule coordinated between the aspartic acids is activated by the abstraction of a proton, enabling the polarized water to attack the carbonyl carbon of the substrate's scissile bond. The tetrahedral transition state formed on nucleophilic attack seems to be uncharged, unlike the oxyanion that arises during catalysis by serine and cysteine proteases.

Metalloproteases (EC 3.4.24; **d** in the figure) use a coordinated metal, often zinc, in their catalytic mechanism. In many soluble metalloproteases, such as thermolysin and the matrix metalloproteases, coordination is accomplished by three histidines, or two histidines and an acidic side chain. A water molecule serving as an additional zinc ligand is hydrogen-bonded to a glutamate, which abstracts a proton from the attacking water molecule. The zinc ion itself stabilizes the oxyanion.

In the image, the pale blue curves in **a** and **b** represent oxyanion holes; the larger curves represent the enzyme schematically. Red arrows indicate movement of electron pairs. Blue dotted lines represent hydrogen bonds or other electrostatic interactions. Grey lines represent the continuation of the substrate polypeptide to either side of the peptide bond that is represented explicitly.

after a *Drosophila melanogaster* mutant[15], are involved in signalling by members of the epidermal growth factor (EGF) family[16,17]. The precursor of the *Drosophila* EGF-domain protein Spitz is trafficked to the Golgi apparatus where it is cleaved by rhomboid protease, and the luminal cleavage product is secreted to activate EGF receptors on other cells (Fig. 1b). Human rhomboids have been shown to cleave cell-surface proteins[18,19], but further studies will be needed to appreciate the full range of their signalling activities. The function of prokaryotic rhomboids is poorly understood, but one enzyme seems to be associated indirectly with intercellular signalling pathways. The rhomboid AarA of the bacterium *Providencia stuartii* cleaves and activates the type I membrane protein TatA, a central component of the twin-arginine translocase, which exports folded proteins across the inner membrane. One of these translocated proteins may be essential for quorum sensing, a mechanism for signalling cell density[20].

Like other I-CLiPs, rhomboids also participate in diverse processes as well as signalling. For example, rhomboids from the malaria parasite *Plasmodium falciparum* and the protozoan *Toxoplasma gondii* cause shedding of parasite adhesins during the invasion of host cells[21]. In a different context, rhomboids process mitochondrial proteins and regulate mitochondrial membrane dynamics. In yeast, a mitochondrial rhomboid cleaves Mgm1, a dynamin-like GTPase that promotes mitochondrial fusion, such that a deficiency of either Mgm1 or the mitochondrial rhomboid results in the fragmentation of mitochondria[22,23]. However, the mammalian mitochondrial rhomboid PARL does not seem to affect mitochondrial morphology but rather controls mitochondria-related apoptosis[24,25]. Amino-acid sequence comparisons have uncovered a widespread group of degenerate rhomboid-like proteins, termed iRhoms for 'inactive rhomboids', which usually lack one or more of the catalytic residues and have a GPX sequence in place of the typical GXS motif containing the catalytic serine[13,14]. The iRhoms perhaps have chaperone or regulatory functions, as the peptide-binding capabilities of proteases may be useful even in the absence of catalytic activity[26,27].

Intramembrane aspartyl proteases

There are two main classes of eukaryotic aspartyl I-CLiP: PSEN and SPP. Both classes have nine transmembrane segments, with the two catalytic aspartates located in a YDX_nLGhGD motif[28], and they show sensitivity to the same inhibitors[29]. Despite their shared active-site motifs, SPP and PSEN differ in certain structural and functional features. SPP functions independently of other proteins, whereas PSEN is the catalytic subunit within γ-secretase[30,31], a multimeric intramembrane complex containing three additional proteins. The apparent stoichiometry of γ-secretase is 1:1:1:1[32,33], but the presence of two copies of PSEN cannot be ruled out[34]. An additional difference is that PSEN undergoes endoproteolytic maturation, perhaps autocatalytically (PSEN cleaves itself)[35]; in contrast, no post-translational cleavage has been observed for SPP. Most significantly, however, PSEN and SPP are thought to have opposite membrane topologies and to cleave oppositely oriented substrate transmembrane segments[36] (Fig. 2).

PSEN, or presenilin, derives its name from its role in the formation of senile plaques in Alzheimer's disease. Specifically, the amyloid protein precursor (APP) is a target of γ-secretase, which cleaves APP to yield the amyloid-β peptide (Aβ) (Fig. 1c). Mutations in PSEN affect the boundaries of the Aβ peptides generated, and thereby the likelihood that they will aggregate and cause disease[37]. Analogous to the requirement for S1P before cleavage by S2P metalloproteases, APP preprocessing is required before cleavage by γ-secretase[37]. Another target of PSEN is Notch[38–40], a determinant of cell differentiation during development and growth. Notch is a single-pass membrane protein found on the cell surface, and its interaction with cognate transmembrane ligand proteins on neighbouring cells induces steps that lead to its cleavage by γ-secretase. The released Notch intracellular domain then enters the nucleus to promote gene expression. Remarkably, functions of PSEN have been suggested that are independent of its activity as the protease subunit within γ-secretase[41], including roles in intracellular trafficking, apoptosis and calcium homeostasis.

The second class of aspartyl I-CLiP is SPP. After the activity of signal peptidase, SPP catalyses intramembrane cleavage of the released signal peptides to clear the membrane[42]. Other physiological functions attributed to SPP are also related to the intramembrane cleavage of signal peptides. For example, products of cleaved signal peptides from class I major histocompatibility complex (MHC) proteins are presented on the cell surface by the human lymphocyte antigen-E (HLA-E) immune receptor, signalling to natural killer (NK) cells that MHC processing is normal[43] (Fig. 1d). SPP also has a role in processing proteins of the heptatitis C virus and is required for virus maturation[44].

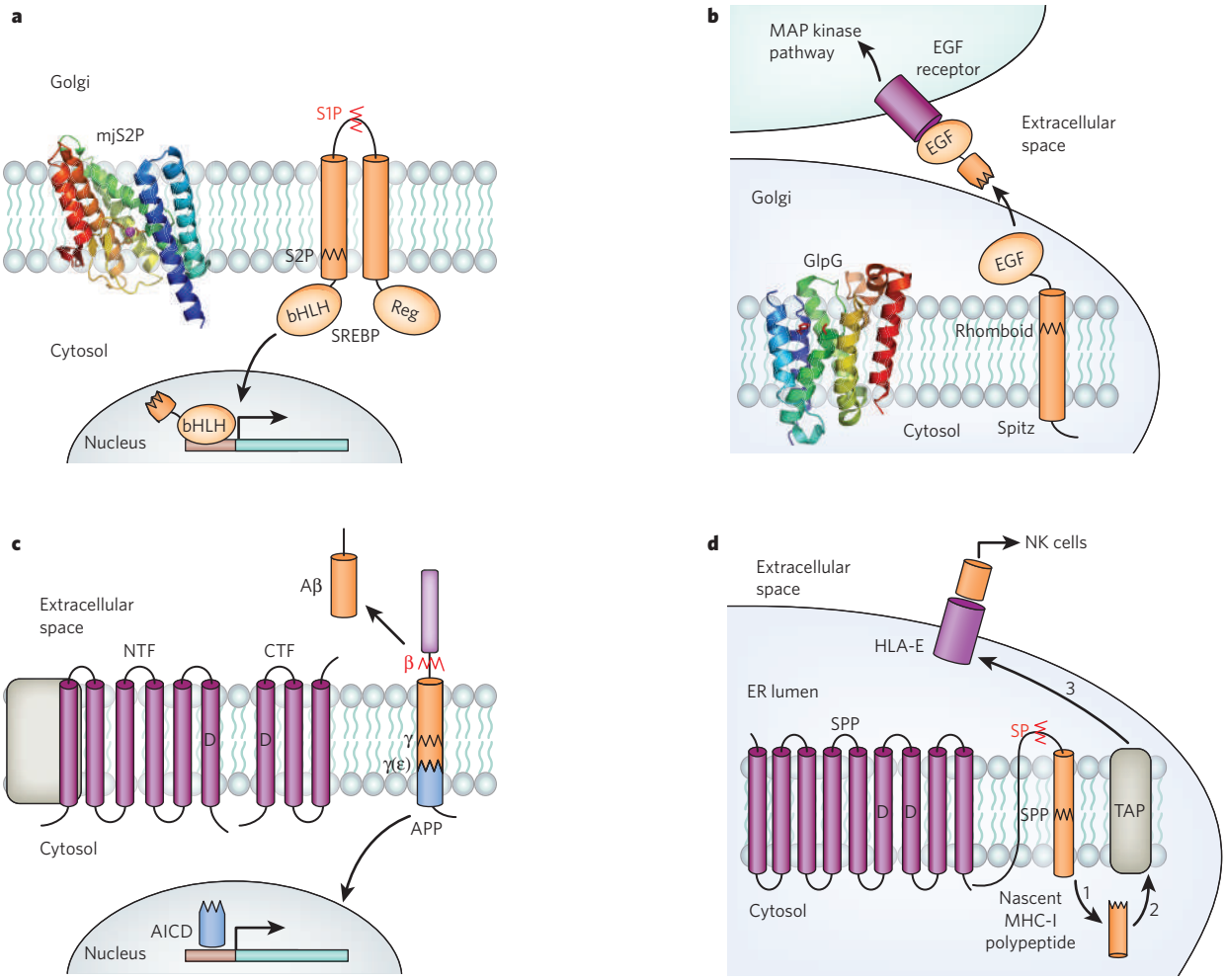


Figure 1 | I-CLiP families and their signalling-related functions. **a**, Transcriptional activation by the cleavage of sterol regulatory element-binding protein (SREBP) by site-2 protease (S2P). When the cellular level of sterols drops, the SREBP precursor protein is transported to the Golgi apparatus, where it is first cleaved by site-1 protease (S1P) (red zigzag line) and then by S2P (black zigzag line). The liberated transcription-factor domain (bHLH) travels to the nucleus and directs the transcription of target genes. mjS2P, crystal structure of the *Methanocaldococcus jannaschii* S2P (ref. 49). **b**, Activation of the epidermal growth factor (EGF) receptor by the rhomboid-dependent release of its ligand in *Drosophila*. The EGF receptor ligand Spitz is transported to the Golgi apparatus, where it is cleaved by rhomboid. The luminal product is secreted and activates EGF receptors on neighbouring cells. GlpG, crystal structure of *Escherichia coli* rhomboid[55]. **c**, Release of amyloid β-peptide (Aβ) by γ-secretase (γ). The catalytic component of γ-secretase, presenilin, is found in complex with three other proteins (nicastrin (NCT), anterior pharynx defective 1 (Aph-1) and presenilin enhancer 2 (Pen-2); complex shown schematically in grey on the left of the figure). Presenilin undergoes endoproteolytic cleavage to create an amino(N)-terminal fragment (NTF) and a carboxy (C)-terminal fragment (CTF). Amyloid precursor protein (APP) on the plasma membrane is first cleaved by β-secretase (red zigzag line and β) at an extracellular domain, and then by γ-secretase (black zigzag line and γ) to release Aβ, which is secreted from cells. The remnant is further cleaved by γ-secretase at the ε site (black zigzag line and γ(ε)) to create APP intracellular domain (AICD), which is thought to have a role in transcription regulation[37]. **d**, Generation of human lymphocyte antigen E (HLA-E) epitopes by signal peptide peptidase (SPP)-catalysed intramembrane cleavage. During biosynthesis, the signal sequence of a major histocompatibility complex class I (MHC-I) molecule is first cleaved by signal peptidase (SP) (red zigzag line) and then by SPP (black zigzag line). The resultant fragment is released to the cytosol (step 1) where it is further cleaved (not shown) and transported back to the lumen of the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) (step 2). The epitope-containing fragment is then loaded onto the HLA-E molecule and transported to the cell surface for presentation to natural killer (NK) cells (step 3). D, aspartates in active site.

In addition to PSEN and SPP, sequences coding for additional putative aspartyl I-CLiPs have been identified in the human genome[45], and their products have been shown in several cases to be functional proteases[46,47]. These proteins are predicted to share the SPP orientation in the membrane and therefore are referred to as SPP-like (SPPL). Whereas SPP and PSEN are found only in plants and animals, expansion of the intramembrane aspartyl protease family to include SPPLs uncovered fungal and archaeal enzymes as well[45].

I-CLiP structure and mechanism

X-ray crystallography of membrane proteins, particularly those of eukaryotic origin, is challenging. Of the roughly 150 unique structures of membrane proteins obtained, those from eukaryotes make up only a small fraction[48]. Therefore, it is useful that I-CLiPs are found in all kingdoms of life. Bacterial and archaeal I-CLiPs can serve as structural models for eukaryotic enzymes, as the core functional features should be present in both. What fundamental features do we expect to find in intramembrane proteases?

Regardless of whether a protease functions in an aqueous environment or in a membrane, its structure must presumably meet several fundamental requirements (Box 1). Mechanisms must be in place to bind the substrate polypeptide and orient the scissile bond, to polarize the nucleophile, to bring about nucleophilic attack on a peptide bond and to stabilize a negatively charged oxyanion intermediate. How can I-CLiPs do all this inside the membrane? How can they overcome the apparent incompatibility between the membrane environment and the requirement for water for hydrolysis of the peptide bond? Our understanding of I-CLiP mechanisms made a dramatic advance when high-resolution crystal structures of several I-CLiPs, from the metal-dependent and serine protease families, were obtained. The more recalcitrant aspartyl I-CLiPs have been investigated using electron microscopy and show promise for future X-ray crystallographic breakthroughs. The structural studies provide a bridge to the well-characterized water-soluble proteases, as the I-CLiP active sites can be compared in molecular detail with enzymes whose mechanisms have been studied for decades. Furthermore, the structures establish a framework in which hypotheses can be made regarding interactions with substrates and with membrane lipids.

S2P mimics water-soluble metalloproteases

The structure of the *Methanocaldococcus jannaschii* mjS2P I-CLiP[49] was obtained after pursuing the technically most tractable of 40 bacterial and archaeal S2P sequences cloned and expressed. Two of its six transmembrane segments provide active-site residues, whereas the other four, as well as a small β-sheet predicted to span only one leaflet of the membrane, support and enclose the active-site region (Fig. 3a). Although a membrane-embedded active site may be a counterintuitive location for a proteolytic event, the mjS2P structure reveals how principles from water-soluble proteases can be transferred to the plane of the membrane. Many soluble metalloproteases, such as thermolysin (Fig. 3a inset) and the matrix metalloproteases, coordinate a metal ion using two histidines on successive turns of an α-helix and a third residue from a nearby helix or loop. This zinc-coordinating motif is clearly applicable to transmembrane helices, as seen in mjS2P using His 54, His 58 and Asp 148 (ref. 49) (Fig. 3a). The glutamate involved in water deprotonation by soluble metalloproteases[50] also has a direct counterpart in mjS2P (Glu 55). The mjS2P asparagine (Asn 140) that may interact with a substrate carbonyl is in a comparable position[51] to asparagines in the metalloproteases carboxypeptidases A and B. In thermolysin, an arginine appearing some 5 Å from the zinc and above the third zinc ligand presumably fulfils a similar function. The proper positioning of Asn 140 in mjS2P relative to the downstream zinc-coordinating residue Asp 148 is accomplished by a break and dislocation of the TM4 helix. Helix-breaking prolines are consistently found between the conserved asparagine and aspartate residues in the TM4 region of S2P-family sequences[8,12].

The mjS2P crystal structure suggests a means by which substrate access to the active sites of S2P enzymes can be controlled. Two conformations of the molecule were trapped in the crystal, and a comparison suggests a gating mechanism in which TM1 and the TM5–TM6 pair slide open to either side of the catalytic centre (Fig. 3b). Features that may enable alternate conformations of mjS2P are the general lack of polar or charge–charge interactions between TM1 and TM2 and hence weak specificity in packing, as well as two possible salt-bridge partners on TM4 for a glutamate on TM6, which may be the basis for a conformational switch. Hinge motions exposing the catalytic cleft of water-soluble metalloproteases have also been observed[52], but the structural variability in mjS2P is more dramatic. How does water, in effect the second substrate of a hydrolytic reaction, access the membrane-embedded mjS2P active site? This problem seems to be solved by a channel perpendicular

Table 1 | I-CLiP families and their major substrates

I-CLiP	Pathway	Substrates	References
S2P	Cholesterol and fatty acid synthesis	SREBP	4, 5
	Unfolded protein response	ATF6	9
	Stress response	RseA	10
	Sporulation	Pro-σ^K	8
	Conjugation	cAD1 precursor	80
Rhomboid	EGF signalling	Spitz, Keren, Gurken	16, 17
	Quorum sensing	TatA	20
	Parasite invasion	Adhesins	21
	Mitochondria dynamic	Mgm1, Ccp1	22, 23
Presenilin	Alzheimer's disease	APP	30, 31, 37
	Notch signalling	Notch	38–40
	Tyrosine kinase receptor	ErbB-4	81
	Cell adhesion	CD44, E-cadherin	82, 83
SPP	HLA-E signalling	MHC-I	43
	Hepatitis C virus maturation	Hepatitis C virus core protein	44
	Regulation of immune system	TNF-α	46, 47
	GVB-B maturation	GVB-B core protein	84

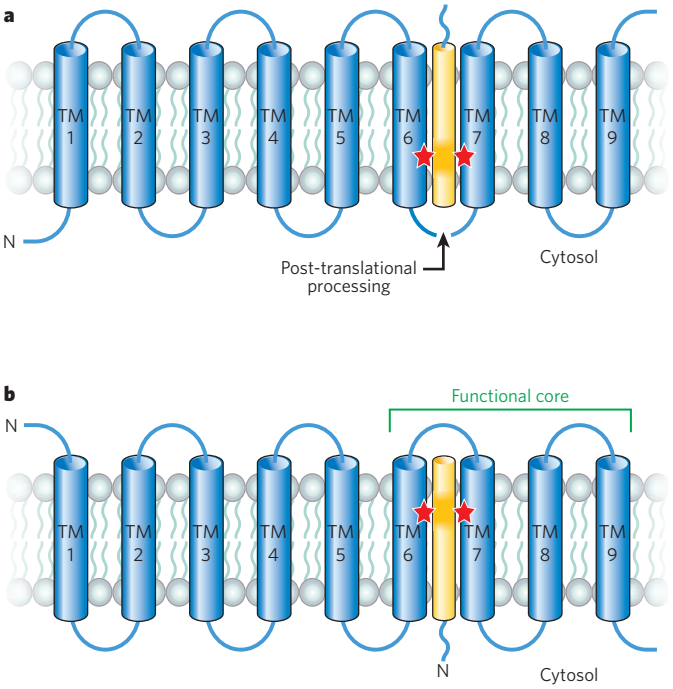


Figure 2 | Secondary structure of presenilin and SPP. Proposed topological organization of presenilin (**a**) and SPP (**b**) and their substrates in the membrane. Blue cylinders are putative transmembrane (TM) helices, and yellow cylinders are transmembrane substrates. Red stars indicate catalytic aspartates. Endoproteolysis takes place between TM6 and TM7 (ref. 35).

to the plane of the membrane that allows water to penetrate from an end of the protein that is exposed to aqueous solution. Although narrow in the closed conformation, the channel may widen during the same gating events that give the substrate polypeptide access to the active site.

The structure of a serine protease active site

Despite pioneering work on serine I-CLiPs in *Drosophila*, a prokaryotic enzyme served as the model system for structural studies. Crystal structures of the bacterial I-CLiP GlpG[53–56] show five transmembrane segments encircling a central helix that spans only part of the membrane (Fig. 4a), leaving a water-filled catalytic cavity exposed to the periplasmic side of the membrane. A helical hairpin between TM1 and TM2 lies perpendicular to the transmembrane helices, embedded in the outer leaflet of the membrane. Structurally, this hairpin supports the back wall of the active-site cavity and may stabilize the GlpG fold. It may also have a role in setting the depth of the protein in the membrane and sensing or affecting the surrounding membrane environment.

The GlpG active-site dyad (Fig. 4a) is formed by a serine at the N terminus of the central helix and a histidine on a juxtaposed parallel transmembrane segment, a framework very different from the classic serine protease trypsin. In addition to the trypsin-like enzymes, other convergently evolved serine protease families are known. Subtilisin and ClpP each have a serine nucleophile at the N terminus of an α-helix, as in GlpG, and the histidine of subtilisin also arises from a helical segment. But the similarity ends there because substrate polypeptides lie across the subtilisin active site in the opposite orientation to that predicted for GlpG (Fig. 4b), based on the topology of the transmembrane segments that GlpG cleaves and the assumption that the substrate approaches the GlpG active site from between TM2 and TM5. In addition to the catalytic dyad, a second class of conserved residue in serine I-CLiPs is the HX_4HX_3N sequence from TM2. Together with a backbone NH group, this motif is in the proper position to constitute a proton-donating pocket, the oxyanion hole (Fig. 4b), which stabilizes an intermediate formed during substrate cleavage. One of the GlpG structures has the phosphate group of a phospholipid in this site[55], indicating that this pocket can accommodate negatively charged groups. A third class of important conserved residue is a set of glycines that allow the close approach of the three catalytically important helices (Fig. 4a). In particular, glycines on the TM4-interacting face of two successive turns of TM6 bring the serine and histidine of the active-site dyad into proximity, as well as being a common stabilizing motif in membrane proteins. A glycine two residues before the catalytic serine may leave room for the putative oxyanion hole. Finally, the glycine-rich segment upstream of the active-site serine in GlpG positions hydrogen-bond donor and acceptor groups towards the cavity above the active-site serine, perhaps to bind the polypeptide backbone N-terminal to the scissile bond (Fig. 4b).

The multiple GlpG structures determined[53–58] provide an opportunity to explore the differences between the models and their implications for the interaction of rhomboid proteases with substrate. As observed for mjS2P, structural variability among the GlpG models suggests a gating mechanism and a route for the substrate to approach the active site. TM5, which packs against the front face of the GlpG helix containing the active-site serine, exhibits the greatest variability in structure and position (Fig. 4c). It may move aside to give the substrate access to the central TM4 segment.

It is increasingly appreciated that membrane proteins do not passively traverse the membrane but rather manipulate the bilayer by displacing lipids from aqueous pockets, and perhaps also by bending, compressing or otherwise deforming the membrane. An example for comparison is the bowl-shaped trimeric glutamate-transporter homologue, which allows a lake of aqueous solution to reach the midpoint of the membrane[59]. It was noted that the hydrophobic belt around the GlpG structure is thinner than expected from the typical thickness of the lipid bilayer, reducing the gap between the active site and the aqueous environment and potentially destabilizing the substrate[58]. Furthermore, the GlpG structure inspires an additional provocative hypothesis for the role of protease–lipid interactions in bringing about substrate recognition and cleavage[60]. Both GlpG and mjS2P have structural protrusions — a helical hairpin in GlpG, and a small, four-stranded β-sheet in mjS2P — at right angles to the transmembrane segments and of the right size to occupy only one of the membrane leaflets. It is possible that these odd projections may influence local lipid structure, which in turn may affect the stability and dynamics of potential substrate transmembrane segments. These structures and hypotheses may well inspire new approaches for exploring membrane protein and lipid conformational dynamics.

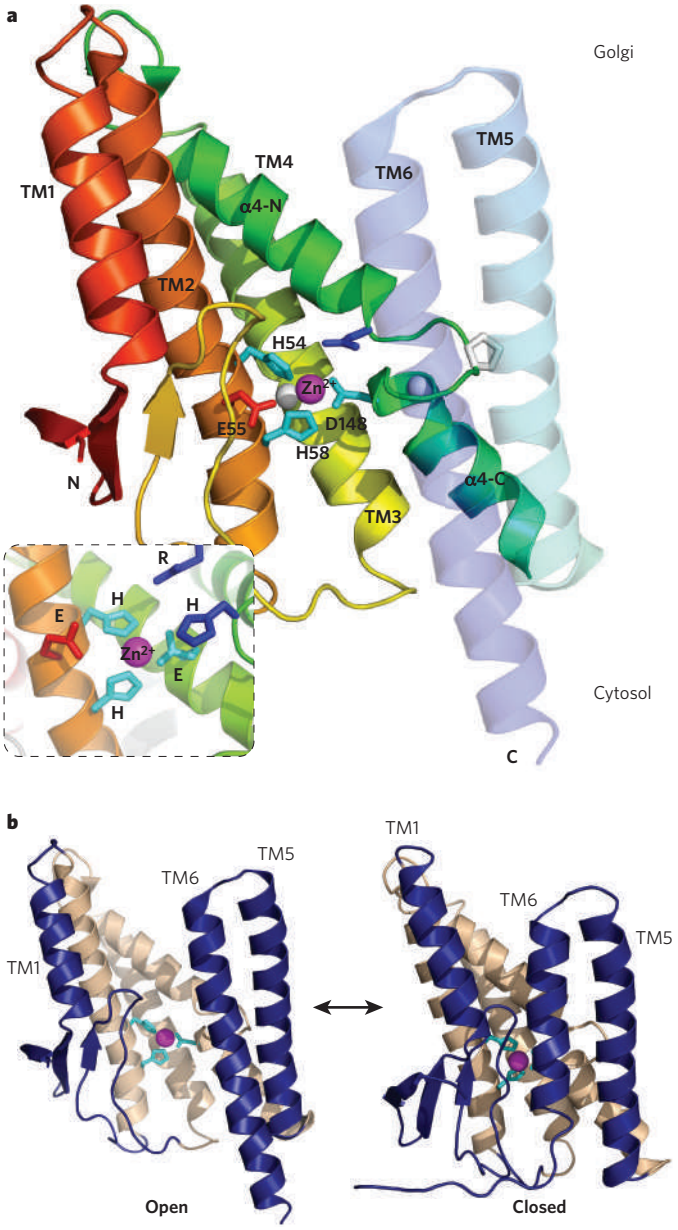


Figure 3 | Structure of *Methanocaldococcus jannaschii* S2P (mjS2P). a, Ribbon diagram of archaebacterial S2P (PDB code, 3B4R; chain A) coloured with a rainbow gradient from N (red) to C (blue) terminus. The bound zinc ion is purple and the side chains of its coordinating residues are shown as cyan sticks. The Cα atoms of conserved glycine residues are shown as white spheres, and a conserved proline as white sticks. The inset shows the active-site region of the water-soluble metalloenzyme thermolysin with side chains of important residues labelled in the one-letter amino-acid code. **b**, The two observed conformations of mjS2P are shown with mobile transmembrane helices in dark blue. The distance between TM1 and the TM5–TM6 pair differs between the open (left) and closed (right) structures.

Progress in determining aspartyl I-CLiP structures

The bilobed monomeric water-soluble aspartyl proteases such as pepsin contain two aspartates, whereas the homodimeric aspartyl proteases such as the retroviral proteases contain only one aspartate per subunit. All known aspartyl I-CLiPs contain two conserved aspartates, so the minimum catalytic unit is probably a monomer. If the metal-dependent and serine I-CLiPs are a precedent, then aspartyl I-CLiPs will mimic soluble aspartyl proteases in the positions of the active-site residues. The helices containing the conserved aspartates are then predicted to be juxtaposed spatially, with the aspartates projecting into a water-filled channel. This model already has some experimental support for PSEN[61], but no crystal structures have yet been reported for aspartyl I-CLiPs.

In response to the different challenges posed by each subfamily, PSEN (γ-secretase) and SPP are taking different routes to structure determination. Successful heterologous expression of SPP, coupled with the surprising finding that SPP has a functional core domain composed of only the four C-terminal transmembrane segments[62] (Fig. 2), should facilitate crystallization and the solution of a high-resolution structure. In contrast, the size and complexity of γ-secretase, which present difficulties for crystallography, are assets for electron microscopy. A 12-Å-resolution cryo-electron-microscopy structure of γ-secretase[33] has revealed the overall shape of the particle. Found to be larger in its membrane region than expected from the size of the packed 19-transmembrane segments of its four subunits, the γ-secretase structure supports the existence of membrane-embedded aqueous pockets and surface grooves. These features may represent entry points for water and substrate to the catalytic centre. However, a better understanding of the γ-secretase architecture will certainly require the positions of the four subunits to be assigned within the reconstructed particle, and insight into the catalytic mechanism requires high-resolution data for at least the PSEN subunit. In addition, determining the structure of PSEN alone may help to clarify intriguing evidence that it has γ-secretase-independent physiological roles, as discussed above.

Selecting I-CLiP substrates

Even I-CLiPs within the same family have sparse sequence similarity and have different biological functions in different organisms, cell types and organelles, so homologous I-CLiPs naturally act on different substrates. Nevertheless, I-CLiPs can in some cases cleave synthetic model substrates or non-cognate substrates from different species[63,64], indicating that either there is minimal substrate specificity or there are shared structural features in the proteases, substrates or both. The only universal feature of the I-CLiP substrates documented to date is that they are integral membrane proteins with a single transmembrane segment. How do I-CLiPs recognize and interact with transmembrane substrates and avoid cleaving other membrane proteins? Some sequence preferences have been observed for individual I-CLiPs[64], but attention has generally been drawn to structural features as a means of substrate selection.

It is widely thought that I-CLiP substrates are particularly susceptible to the unwinding of transmembrane helices. Cleavage by *Drosophila* Rhomboid-1 requires a putative helix-destabilizing GA amino-acid sequence in its substrate Spitz, and insertion of the GA motif into a non-substrate transmembrane segment is sufficient to convert it into an efficient Rhomboid-1 substrate[65]. Helix-breaking residues are also required for the cleavage of signal peptides and SREBP by SPP and S2P, respectively[42,66], and regions C-terminal to the cleavage sites of γ-secretase substrates have been predicted to exhibit decreased helicity[67]. Just as secondary and tertiary structures mask potential cleavage sites in substrates of water-soluble proteases, transmembrane helicity is expected to shield peptide bonds from I-CLiP activity. The disruption of local secondary structure is therefore presumably necessary for intimate interaction between I-CLiPs and the substrate backbone. However, no I-CLiP substrate has been crystallized, either alone or with its protease. In addition, there are other examples of substrates that lack obvious helix-breaking residues in the scissile-bond region[19,68]. In such cases, and perhaps in general, the I-CLiPs may themselves contribute to substrate unwinding either directly, through the stabilization of non-helical configurations, or indirectly, through effects on lipid structure and dynamics, as described above.

Perhaps not surprisingly, the depth of the scissile bond within a substrate transmembrane region generally correlates with the depth of the I-CLiP active site. For example, the S2P active site is located close to the cytosol, and substrates are cleaved in the cytosolic half of their transmembrane segments[69]. For comparison, the serine I-CLiP active site is located distal to the cytosol, and cleavage occurs near, or at the outer edge of, the substrate transmembrane segment[22,70,71]. Cleavage sites just outside the predicted transmembrane region are consistent with a

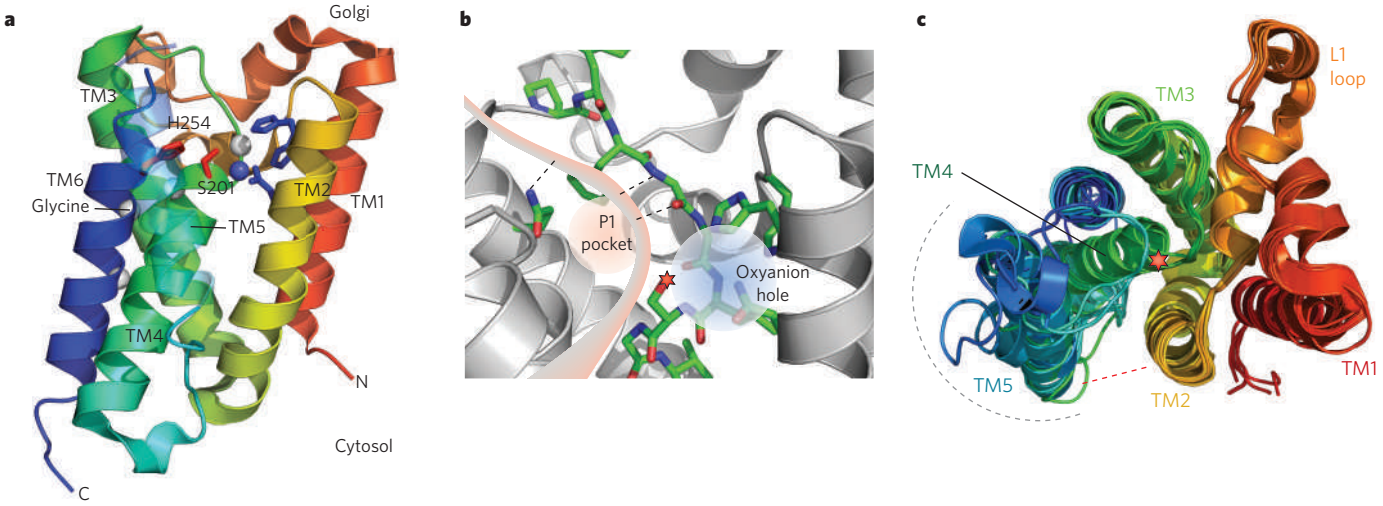


Figure 4 | Structure of *Escherichia coli* GlpG. a, Ribbon diagram of bacterial GlpG (PDB code, 2irv; chain A) coloured with a rainbow gradient from N (red) to C (blue) terminus. The side chains of the active-site Ser/His dyad are shown as red sticks. Side chains potentially contributing to the oxyanion hole are shown as blue sticks, and a backbone NH group that may participate is shown as a blue sphere. Conserved glycines are shown as white spheres. **b**, A model for how substrates may lie across the active site of GlpG, with approximate positions indicated for functionally important sites in the enzyme (the oxyanion hole and the P1 pocket for accommodating the side chain of the residue just before the cleavage site). Dashed lines represent potential hydrogen bonds between enzyme and substrate. The catalytic serine residue is indicated by a red star. **c**, All the GlpG structures determined to date (PDB codes 2NR9, 2IRV, 2NRF, 2O7L, 2IC8 and 3B44) are superposed to illustrate the flexibility of transmembrane helix 5. Structures are coloured as in **a** and viewed from above (from the periplasm). The catalytic serine residue is indicated by a red star. The red double-headed arrow indicates the variable distance between TM5 and TM2.

mechanism in which a non-helical segment loops into the active-site cavity from the aqueous phase. In addition to depth, the direction of the substrate polypeptide also correlates with the orientation of the I-CLiP in the membrane. I-CLiPs of the same class that have opposite orientations in the membrane cleave substrates with opposite orientations (Fig. 2). In other words, the polarity of the substrate polypeptide chain as it lies across the active site seems to be conserved in each I-CLiP family, with one possible exception[72].

Many I-CLiPs only select substrates once they have been precleaved by other proteases, which provide a regulatory step in the process. Understanding the interactions between I-CLiPs and substrates requires a physical explanation for the role of precleavage. Perhaps the shedding of soluble domains exposes a substrate determinant that is required for interaction with I-CLiPs[73]. Alternatively, precleavage may remove a suppressor domain from the substrate[74] or uncouple transmembrane segments to isolate and expose a single-pass substrate segment. As mentioned, rhomboids do not require precleavage of their substrates, raising the question of how these I-CLiPs are regulated. One recurrent theme in regulating interactions between I-CLiPs and their substrates is localization. Cleavage is prevented when the substrates and enzymes are in different compartments; they must be brought together for hydrolysis to occur. Such a mechanism for the control of cleavage has been observed for rhomboid proteases[17,72], γ-secretase[67] and S2P[75].

Perspectives

The first images of intramembrane proteases provided treasure maps that can be used to explore the biochemical mechanisms behind the functions of I-CLiPs. However, interpreting these images is a major challenge. Independent crystallization experiments yield slightly different structures, and proteins can assume different conformations even within a single asymmetric unit, as observed for mjS2P[49] and some of the GlpG studies[53–56]. Do the observations of multiple conformations and apparent flexibility merely flag the more dynamic regions of the protein, or do they accurately represent the range of motions accessible to and assumed by the protein in its native membrane environment? For example, four turns of the C-terminal helix are unwound in one of the mjS2P metalloprotease structures[49]. Such a conformation would presumably not be favoured in a membrane, where intrahelical hydrogen bonds are stabilized. It is rather subjective to ascribe this conformational difference to a crystallization artefact while concluding that the conformational differences suggestive of a sliding gate are functionally significant. Mutagenesis experiments go some way to addressing the significance of the flexibility observed in crystallographic experiments[71], but new methods to analyse intramembrane protease dynamics *in situ* would be most welcome.

Structures of I-CLiP complexes with inhibitors or substrate analogues may help focus speculation regarding the interactions between enzyme and substrate. It might seem that the problems of membrane-protein crystallography will be compounded by the increased complexity of adding another component, but inhibitors may actually help to stabilize I-CLiPs, making crystallization easier. A great deal of effort has been devoted to defining consensus sequences for substrate recognition and cleavage by I-CLiPs, a task that has met with serious difficulties[76]. Most soluble proteases lack strict sequence specificity, so I-CLiP willingness to interact with various substrates is not unique and should not be considered a barrier to future progress. Attention should then turn to finding peptide-based inhibitors of crystallizable I-CLiPs, as information on the interactions between soluble proteases and substrates has largely been extrapolated from enzyme–inhibitor complexes. The mode of interaction between I-CLiPs and their substrates is likely to be conserved through evolution, so structures of complexes containing prokaryotic enzymes would be informative.

In addition to helping focus research on I-CLiP mechanisms, acknowledging the promiscuity of I-CLiP substrate recognition[67,77] can guide speculation about the origins of these enzymes. The apparent promiscuity evoked the hypothesis that some I-CLiPs may have a role in intramembrane quality-control pathways, either as proteases that cleave damaged membrane proteins[78] or as chaperones that stabilize and regulate trafficking or the function of other membrane proteins[41]. Clearly, the many unique and specific functions of individual I-CLiPs seem to defy generalization. Nevertheless, weak substrate sequence specificity and a broad role in aiding recovery from membrane threats (such as the build-up of misfolded membrane proteins) and insults (foreign toxins) in the deep past may have provided the starting point for the diversification of I-CLiPs and their entry into numerous functional niches. A hypothetical role in clearing the primordial membrane raises the question of whether I-CLiPs can cleave polytopic membrane proteins, which predominate the membrane proteome in prokaryotes[79]. It seems likely that membrane proteins are inaccessible to I-CLiPs if they are tightly folded or associate with partners. However, damaged or misfolded multispanning membrane proteins might expose potentially cleavable transmembrane segments. It remains to be seen whether misfolded multispanning membrane proteins are targets of today's I-CLiPs, either in prokaryotes or eukaryotes. ■

1. Dodson, G. & Chothia, C. Fifty years of pepsin crystals. *Nature* **309,** 309 (1984).
2. Protein Data Bank Newsletter <ftp://ftp.wwpdb.org/pub/pdb/doc/newsletters/bnl/news01_sep74.pdf> (1974).
3. Southan, C. A genomic perspective on human proteases as drug targets. *Drug Discov. Today* **6,** 681–688 (2001).
4. Rawson, R. B. *et al.* Complementation cloning of S2P, a gene encoding a putative metalloprotease required for intramembrane cleavage of SREBPs. *Mol. Cell* **1,** 47–57 (1997).
This paper presents the first identification of an I-CLiP.
5. Brown, M. S., Ye, J., Rawson, R. B. & Goldstein, J. L. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell* **100,** 391–398 (2000).
6. Urban, S. & Freeman, M. Intramembrane proteolysis controls diverse signalling pathways throughout evolution. *Curr. Opin. Genet. Dev.* **12,** 512–518 (2002).
7. Lewis, A. P. & Thomas, P. J. A novel clan of zinc metallopeptidases with possible intramembrane cleavage properties. *Protein Sci.* **8,** 439–442 (1999).
8. Rudner, D. Z., Fawcett, P. & Losick, R. A family of membrane-embedded metalloproteases involved in regulated proteolysis of membrane-associated transcription factors. *Proc. Natl Acad. Sci. USA* **96,** 14765–14770 (1999).
9. Ye, J. *et al.* ER stress induces cleavage of membrane-bound ATF6 by the same proteases that process SREBPs. *Mol. Cell* **6,** 1355–1364 (2000).
10. Alba, B. M. & Gross, C. A. Regulation of the Escherichia coli sigma-dependent envelope stress response. *Mol. Microbiol.* **52,** 613–619 (2004).
11. Ellermeier, C. D. & Losick, R. Evidence for a novel protease governing regulated intramembrane proteolysis and resistance to antimicrobial peptides in *Bacillus subtilis*. *Genes Dev.* **20,** 1911–1922 (2006).
12. Kinch, L. N., Ginalski, K. & Grishin, N. V. Site-2 protease regulated intramembrane proteolysis: sequence homologs suggest an ancient signaling cascade. *Protein Sci.* **15,** 84–93 (2006).
13. Koonin, E. V. *et al.* The rhomboids: a nearly ubiquitous family of intramembrane serine proteases that probably evolved by multiple ancient horizontal gene transfers. *Genome Biol.* **4,** R19 (2003).
14. Lemberg, M. K. & Freeman, M. Functional and evolutionary implications of enhanced genomic analysis of rhomboid intramembrane proteases. *Genome Res.* **17,** 1634–1646 (2007).
15. Mayer, U. & Nüsslein-Volhard, C. A group of genes required for pattern formation in the ventral ectoderm of the *Drosophila* embryo. *Genes Dev.* **2,** 1496–1511 (1988).
16. Urban, S., Lee, J. R. & Freeman, M. *Drosophila* rhomboid-1 defines a family of putative intramembrane serine proteases. *Cell* **107,** 173–182 (2001).
This paper describes the first demonstration of cleavage activity by a rhomboid protease.
17. Lee, J. R., Urban, S., Garvey, C. F. & Freeman, M. Regulated intracellular ligand transport and proteolysis control EGF signal activation in *Drosophila*. *Cell* **107,** 161–171 (2001).
18. Pascall, J. C. & Brown, K. D. Intramembrane cleavage of ephrinB3 by the human rhomboid family protease, RHBDL2. *Biochem. Biophys. Res. Commun.* **317,** 244–252 (2004).
19. Lohi, O., Urban, S. & Freeman, M. Diverse substrate recognition mechanisms for rhomboids: thrombomodulin is cleaved by mammalian rhomboids. *Curr. Biol.* **14,** 236–241 (2004).
20. Stevenson, L. G. *et al.* Rhomboid protease AarA mediates quorum-sensing in *Providencia stuartii* by activating TatA of the twin-arginine translocase. *Proc. Natl Acad. Sci. USA* **104,** 1003–1008 (2007).
21. Dowse, T. J., Koussis, K., Blackman, M. J. & Soldati-Favre, D. Roles of proteases during invasion and egress by *Plasmodium* and *Toxoplasma*. *Subcell. Biochem.* **47,** 121–139 (2008).
22. Herlan, M., Vogel, F., Bornhovd, C., Neupert, W. & Reichert, A. S. Processing of Mgm1 by the rhomboid-type protease Pcp1 is required for maintenance of mitochondrial morphology and of mitochondrial DNA. *J. Biol. Chem.* **278,** 27781–27788 (2003).
23. McQuibban, G. A., Saurya, S. & Freeman, M. Mitochondrial membrane remodelling regulated by a conserved rhomboid protease. *Nature* **423,** 537–541 (2003).
24. Cipolat, S. *et al.* Mitochondrial rhomboid PARL regulates cytochrome c release during apoptosis via OPA1-dependent cristae remodelling. *Cell* **126,** 163–175 (2006).
25. Chao, J. R. *et al.* Hax1-mediated processing of HtrA2 by Parl allows survival of lymphocytes and neurons. *Nature* **452,** 98–102 (2008).
26. Hulko, M., Lupas, A. N. & Martin, J. Inherent chaperone-like activity of aspartic proteases reveals a distant evolutionary relation to double-ψ barrel domains of AAA-ATPases. *Protein Sci.* **16,** 644–653 (2007).
27. Spiess, C., Beil, A. & Ehrmann, M. A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* **97,** 339–347 (1999).

28. Weihofen, A., Binns, K., Lemberg, M. K., Ashman, K. & Martoglio, B. Identification of signal peptide peptidase, a presenilin-type aspartic protease. *Science* **296,** 2215–2218 (2002).
Human SPP was isolated by affinity purification and the protein identified in this paper.
29. Sato, T. *et al.* Signal peptide peptidase: biochemical properties and modulation by nonsteroidal antiinflammatory drugs. *Biochemistry* **45,** 8649–8656 (2006).
30. De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391,** 387–390 (1998).
This paper links intramembrane proteolysis and Alzheimer's disease.
31. Wolfe, M. S. *et al.* Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity. *Nature* **398,** 513–517 (1999).
This paper shows that the transmembrane aspartates in presenilin are critical for its function.
32. Sato, T. *et al.* Active γ-secretase complexes contain only one of each component. *J. Biol. Chem.* **282,** 33985–33993 (2007).
33. Osenkowski, P. *et al.* Cryoelectron microscopy structure of purified γ-secretase at 12 Å resolution. *J. Mol. Biol.* **385,** 642–652 (2009).
34. Schroeter, E. H. *et al.* A presenilin dimer at the core of the γ-secretase enzyme: insights from parallel analysis of Notch 1 and APP proteolysis. *Proc. Natl Acad. Sci. USA* **100,** 13075–13080 (2003).
35. Tolia, A. & De Strooper, B. Structure and function of γ-secretase. *Semin. Cell Dev. Biol.* **20,** 211–218 (2009).
36. Friedmann, E. *et al.* Consensus analysis of signal peptide peptidase and homologous human aspartic proteases reveals opposite topology of catalytic domains compared with presenilins. *J. Biol. Chem.* **279,** 50790–50798 (2004).
37. Steiner, H., Fluhrer, R. & Haass, C. Intramembrane proteolysis by γ-secretase. *J. Biol. Chem.* **283,** 29627–29631 (2008).
38. De Strooper, B. *et al.* A presenilin-1-dependent gamma-secretase-like protease mediates release of Notch intracellular domain. *Nature* **398,** 518–522 (1999).
39. Struhl, G. & Greenwald, I. Presenilin is required for activity and nuclear access of Notch in *Drosophila*. *Nature* **398,** 522–525 (1999).
40. Ye, Y., Lukinova, N. & Fortini, M. E. Neurogenic phenotypes and altered Notch processing in *Drosophila Presenilin* mutants. *Nature* **398,** 525–529 (1999).
41. Hass, M. R., Sato, C., Kopan, R. & Zhao, G. Presenilin: RIP and beyond. *Semin. Cell Dev. Biol.* **20,** 201–210 (2009).
42. Lemberg, M. K. & Martoglio, B. Requirements for signal peptide peptidase-catalyzed intramembrane proteolysis. *Mol. Cell* **10,** 735–744 (2002).
43. Lemberg, M. K., Bland, F. A., Weihofen, A., Braud, V. M. & Martoglio, B. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. *J. Immunol.* **167,** 6441–6446 (2001).
44. McLauchlan, J., Lemberg, M. K., Hope, G. & Martoglio, B. Intramembrane proteolysis promotes trafficking of hepatitis C virus core protein to lipid droplets. *EMBO J.* **21,** 3980–3908 (2002).
45. Ponting, C. P. *et al.* Identification of a novel family of presenilin homologues. *Hum. Mol. Genet.* **11,** 1037–1044 (2002).
46. Friedmann, E. *et al.* SPPL2a and SPPL2b promote intramembrane proteolysis of TNF-α in activated dendritic cells to trigger IL-12 production. *Nature Cell Biol.* **8,** 843–848 (2006).
47. Fluhrer, R. *et al.* A γ-secretase-like intramembrane cleavage of TNF-α by the GxGD aspartyl protease SPPL2b. *Nature Cell Biol.* **8,** 894–896 (2006).
48. Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **18,** 581–586 (2008).
49. Feng, L. *et al.* Structure of a site-2 protease family intramembrane metalloprotease. *Science* **318,** 1608–1612 (2007).
This paper reports the first crystal structure of a metalloprotease.
50. Matthews, B. W. Structural basis of the action of thermolysin and related zinc peptidases. *Acc. Chem. Res.* **21,** 333–340 (1988).
51. Christianson, D. W. & Lipscomb, W. N. Carboxypeptidase A. *Acc. Chem. Res.* **22,** 62–69 (1989).
52. Hausrath, A. C. & Matthews, B. W. Thermolysin in the absence of substrate has an open conformation. *Acta Crystallogr. D* **58,** 1002–1007 (2002).
53. Wang, Y., Zhang, Y. & Ha, Y. Crystal structure of a rhomboid family intramembrane protease. *Nature* **444,** 179–180 (2006).
This paper reports the first crystal structure of the intramembrane serine protease GlpG.
54. Wu, Z. *et al.* Structural analysis of a rhomboid family intramembrane protease reveals a gating mechanism for substrate entry. *Nature Struct. Mol. Biol.* **13,** 1084–1091 (2006).
55. Ben-Shem, A., Fass, D. & Bibi, E. Structural basis for intramembrane proteolysis by rhomboid serine proteases. *Proc. Natl Acad. Sci. USA* **104,** 462–466 (2007).
This paper describes the crystal structure of GlpG in two different conformations.
56. Lemieux, M. J., Fischer, S. J., Cherney, M. M., Bateman, K. S. & James, M. N. The crystal structure of the rhomboid peptidase from *Haemophilus influenzae* provides insight into intramembrane proteolysis. *Proc. Natl Acad. Sci. USA* **104,** 750–754 (2007).
57. Wang, Y. & Ha, Y. Open-cap conformation of intramembrane protease GlpG. *Proc. Natl Acad. Sci. USA* **104,** 2098–2102 (2007).
58. Wang, Y., Maegawa, S., Akiyama, Y. & Ha, Y. The role of L1 loop in the mechanism of rhomboid intramembrane protease GlpG. *J. Mol. Biol.* **374,** 1104–1113 (2007).
59. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
60. Bondar, A. N., del Val, C. & White, S. H. Rhomboid protease dynamics and lipid interactions. *Structure* **17,** 395–405 (2009).
61. Tolia, A., Chávez-Gutiérrez, L. & De Strooper, B. Contribution of presenilin transmembrane domains 6 and 7 to a water-containing cavity in the γ-secretase complex. *J. Biol. Chem.* **281,** 27633–27642 (2006).
62. Narayanan, S., Sato, T. & Wolfe, M. S. A C-terminal region of signal peptide peptidase defines a functional domain for intramembrane aspartic protease catalysis. *J. Biol. Chem.* **282,** 20172–20179 (2007).
63. Urban, S., Schlieper, D. & Freeman, M. Conservation of intramembrane proteolytic activity and substrate specificity in prokaryotic and eukaryotic rhomboids. *Curr. Biol.* **12,** 1507–1512 (2002).
64. Akiyama, Y. & Maegawa, S. Sequence features of substrates required for cleavage by GlpG, an *Escherichia coli* rhomboid protease. *Mol. Microbiol.* **64,** 1028–1037 (2007).
65. Urban, S. & Freeman, M. Substrate specificity of rhomboid intramembrane proteases is governed by helix-breaking residues in the substrate transmembrane domain. *Mol. Cell* **11,** 1425–1434 (2003).
66. Ye, J., Dave, U. P., Grishin, N. V., Goldstein, J. L. & Brown, M. S. Asparagine-proline sequence within membrane-spanning segment of SREBP triggers intramembrane cleavage by site-2 protease. *Proc. Natl Acad. Sci. USA* **97,** 5123–5128 (2000).
67. Beel, A. J. & Sanders, C. R. Substrate specificity of gamma-secretase and other intramembrane proteases. *Cell. Mol. Life Sci.* **65,** 1311–1334 (2008).
68. Lichtenthaler, S. F. *et al.* Mechanism of the cleavage specificity of Alzheimer's disease γ-secretase identified by phenylalanine-scanning mutagenesis of the transmembrane domain of the amyloid precursor protein. *Proc. Natl Acad. Sci. USA* **96,** 3053–3058 (1999).
69. Duncan, E. A., Dave, U. P., Sakai, J., Goldstein, J. L. & Brown, M. S. Second-site cleavage in sterol regulatory element-binding protein occurs at transmembrane junction as determined by cysteine panning. *J. Biol. Chem.* **273,** 17801–17809 (1998).
70. Maegawa, S., Ito, K. & Akiyama, Y. Proteolytic action of GlpG, a rhomboid protease in the *Escherichia coli* cytoplasmic membrane. *Biochemistry* **44,** 13543–13552 (2005).
71. Baker, R. P., Young, K., Feng, L., Shi, Y. & Urban, S. Enzymatic analysis of a rhomboid intramembrane protease implicates transmembrane helix 5 as the lateral substrate gate. *Proc. Natl Acad. Sci. USA* **104,** 8257–8262 (2007).
72. Tsruya, R. *et al.* Rhomboid cleaves Star to regulate the levels of secreted Spitz. *EMBO J.* **26,** 1211–1220 (2007).
73. Schroeter, E. H., Kisslinger, J. A. & Kopan, R. Notch-1 signalling requires ligand-induced proteolytic release of intracellular domain. *Nature* **393,** 382–386 (1998).
74. Shen, J. & Prywes, R. Dependence of site-2 protease cleavage of ATF6 on prior site-1 protease digestion is determined by the size of the luminal domain of ATF6. *J. Biol. Chem.* **279,** 43046–43051 (2004).
75. Goldstein, J. L., Rawson, R. B. & Brown, M. S. Mutant mammalian cells as tools to delineate the sterol regulatory element-binding protein pathway for feedback regulation of lipid synthesis. *Arch. Biochem. Biophys.* **397,** 139–148 (2002).
76. Freeman, M. Rhomboids: 7 years of a new protease family. *Semin. Cell Dev. Biol.* **20,** 231–239 (2009).
77. Weihofen, A. & Martoglio, B. Intramembrane-cleaving proteases: controlled liberation of proteins and bioactive peptides. *Trends Cell Biol.* **13,** 71–78 (2003).
78. Akiyama, Y., Kanehara, K. & Ito, K. RseP (YaeL), an *Escherichia coli* RIP protease, cleaves transmembrane sequences. *EMBO J.* **23,** 4434–4442 (2004).
79. Daley, D. O. *et al.* Global topology analysis of the Escherichia coli inner membrane proteome. *Science* **308,** 1321–1323 (2005).
80. An, F. Y., Sulavik, M. C. & Clewell, D. B. Identification and characterization of a determinant (eep) on the *Enterococcus faecalis* chromosome that is involved in production of the peptide sex pheromone cAD1. *J. Bacteriol.* **181,** 5915–5921 (1999).
81. Ni, C. Y., Murphy, M. P., Golde, T. E. & Carpenter, G. γ -Secretase cleavage and nuclear localization of ErbB-4 receptor tyrosine kinase. *Science* **294,** 2179–2181 (2001).
82. Okamoto, I. *et al.* Proteolytic release of CD44 intracellular domain and its role in the CD44 signaling pathway. *J. Cell Biol.* **155,** 755–762 (2001).
83. Marambaud, P. *et al.* A presenilin-1/γ-secretase cleavage releases the E-cadherin intracellular domain and regulates disassembly of adherens junctions. *EMBO J.* **21,** 1948–1956 (2002).
84. Targett-Adams, P. *et al.* Signal peptide peptidase cleavage of GB virus B core protein is required for productive infection in vivo. *J. Biol. Chem.* **281,** 29221–29227 (2006).

Acknowledgements E.B. is supported by the Israel Science Foundation and the Yale-Weizmann Collaborative Program. D.F. acknowledges support from the Kimmelman Center for Macromolecular Assemblies.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to D.F. (deborah.fass@weizmann.ac.il) or E.B. (e.bibi@weizmann.ac.il).

Emerging roles for lipids in shaping membrane-protein function

Rob Phillips[1], Tristan Ursell[1], Paul Wiggins[2] & Pierre Sens[3]

Studies of membrane proteins have revealed a direct link between the lipid environment and the structure and function of some of these proteins. Although some of these effects involve specific chemical interactions between lipids and protein residues, many can be understood in terms of protein-induced perturbations to the membrane shape. The free-energy cost of such perturbations can be estimated quantitatively, and measurements of channel gating in model systems of membrane proteins with their lipid partners are now confirming predictions of simple models.

Quantitative analysis is changing the face of biology. An area in which it has provided particularly useful insights is the analysis of the function of membrane proteins, specifically with respect to their interactions with the surrounding lipid molecules. Models and experiments show that rather than being a passive bystander in the function of membrane-bound proteins, the membrane can at times have an essential role in determining the function of these proteins.

Cell membranes are the barriers that separate the cytoplasm of the cell from the external world and internally compartmentalize eukaryotic cells into organelles. Far from being inert, biological membranes are key components in sensory and signalling pathways. They are highly controlled barriers that allow the directed flux of molecules into and out of the cytoplasm, and they have an analogous role in intracellular trafficking and energy production in cellular organelles.

At the microscopic scale, biological membranes are a crowded mix of membrane proteins and their lipid partners. Our understanding of this complicated environment is constantly being refined by new experiments[1]. The data that emerge often reveal functional and quantitative relations between biologically interesting parameters (for example, the open probability for ion channels as a function of driving forces such as voltage or membrane tension), and carry with them an imperative for models of the underlying phenomena. Each generation of new experiments refines the models used to describe membranes, a topic elegantly reviewed elsewhere[1].

One case study that illustrates this interplay between quantitative models and experiments concerns the analysis of the structure and function of mechanosensitive channels[2], reconstituted in simple lipid bilayers[3]. The data demonstrate that the physicochemical properties of the surrounding lipid bilayer result in predictable and stereotyped consequences for channel function in artificial lipid membranes, although the interactions between lipids and membrane proteins have broader significance[4–6]. The concepts we present in this Review are predicted to have functional and structural significance for any protein whose function requires the remodelling of the protein–membrane interface. These models are used first to examine the properties of an isolated channel, followed by examples of an added layer of complexity resulting from membrane-mediated interactions between the channels.

Mechanosensitive channels and biological membranes

The idea that sequence dictates structure, which in turn dictates function, is a second central dogma of biology[7]. A powerful example of this dictum is in the context of membrane proteins. The stunning structures obtained of membrane machines, from the light-gathering apparatus of photosynthesis to the voltage-gated channels that allow neurons to propagate electrical impulses and the bacterial sensors that detect osmotic stress, provide key insights into the mechanisms by which these proteins respond to stimuli such as light, voltage and membrane tension. In many cases, complementary functional studies show us that the lipid bilayer is not a passive bystander in membrane protein function, as shown systematically elsewhere[4]. In this context, the word 'structure' usually refers to atomic positions, but a more coarse-grained picture of structure, captured by ideas from continuum elasticity, can reproduce many important membrane properties. In these models, 'structure' refers to quantities such as the local thickness and curvature of the lipid bilayer surrounding the membrane protein of interest. This is in contrast to molecular dynamics, which explicitly represents the position of every atom of both the protein and the lipid bilayer.

As a concrete example, we will consider bacterial mechanosensitive channels. The structure, function and physiology of mechanosensitive channels have been studied extensively. As shown in Fig. 1, bacterial mechanosensitive channels are gated by membrane tension[3]. More precisely, a pipette is used to grab a patch of membrane containing these channels and the current passing through the protein-encumbered membrane is measured as a function of the pipette suction pressure or membrane tension. These experiments demonstrate a relation between the open probability of the channel and the pipette pressure that is dependent on the properties of the lipid membrane in which the proteins find themselves (such as the tail lengths of the lipids, which can result in a mismatch between the protein and the bilayer thickness)[3,8]. Analysis of the gating free energy reveals that the quantitative dependence of the gating tension on the length of the lipid acyl tail matches the prediction from elastic bilayer models. Studies of mechanosensitive channels therefore reveal not only the importance of the lipid environment but show that, at least in the case of hydrophobic mismatch, where the hydrophobic core of the bilayer has a different thickness from the hydrophobic region of a transmembrane protein, the mechanism can be understood in terms of a coarse-grained elastic model of the bilayer.

[1]Department of Applied Physics, California Institute of Technology, Pasadena, California 91125, USA. [2]Whitehead Institute of Biomedical Research, Cambridge, Massachusetts 02142, USA. [3]UMR Gulliver CNRS-ESPCI, Paris 7083, France.

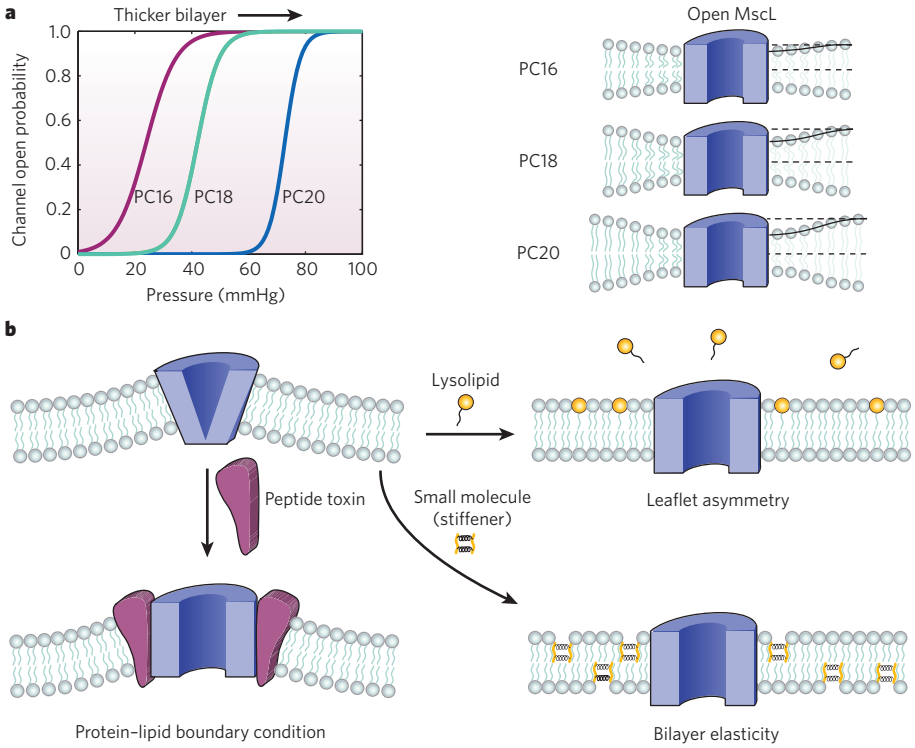


Figure 1 | Ion-channel function and membrane properties. **a**, Ion-channel open probability as a function of pipette pressure for mechanosensitive channels in lipids with different tail lengths. The data are adapted from ref. 3. The curves are an empirical fit to patch-clamp data using the functional form $p_{open} = 1/(1 + \exp(-\alpha(P-P_{1/2}))$, with the parameters α and $P_{1/2}$ as fitting parameters. The diagrams on the right show how different tail lengths imply a different hydrophobic mismatch as a result of the boundary conditions at the protein–lipid interface. MscL, mechanosensitive channel of large conductance. PC16, PC18 and PC20 are phospholipid bilayers with lipids with acyl chain lengths of 16, 18 and 20 carbons, respectively. **b**, Membrane doping and membrane protein function. The diagrams show hypothetical mechanisms whereby the insertion of various molecules can alter the protein–membrane interaction. For example, the asymmetrical insertion of lysolipids in the membrane produces a torque on the protein. The introduction of toxins can alter the boundary conditions between the protein and the surrounding lipids. Finally, small molecules can stiffen the membrane. In principle, all these effects could alter the gating characteristics of a channel.

A second example of the influence of the lipid environment on the function of membrane proteins is provided by the effects of membrane doping (by toxins, lipids or cholesterol) on channel activity (Fig. 1b). Certain lipid species and other membrane components are clearly required for proper protein function[9,10], but studies using toxins support the idea that the membrane is also a generic mechanical medium with which proteins interact. Rather than having evolved to target a specific channel, some toxins impair the function of multiple membrane proteins, and some small molecules, such as capsaicin[11], and peptide toxins, like those found in spider venom[12], target membrane channels across many species. These broad-ranging effects favour a mechanism that targets a generic property of membrane proteins. It has therefore been proposed that these toxins affect the interactions with the membrane itself. But can these toxins be understood in terms of a coarse-grained membrane model?

Many studies have shown that bilayer thickness, bending stiffness and monolayer spontaneous curvature can affect the function of embedded proteins[4,13]. Indeed, although the role of certain proteins (such as mechanosensitive channels) is to respond to membrane mechanical stress, in principle this stress can alter the function of any membrane protein. For example, the dimerization kinetics of the channel-forming peptide gramicidin A can be controlled by an externally applied mechanical stress on the membrane, resulting in membrane thinning and decreasing the hydrophobic mismatch between the membrane and

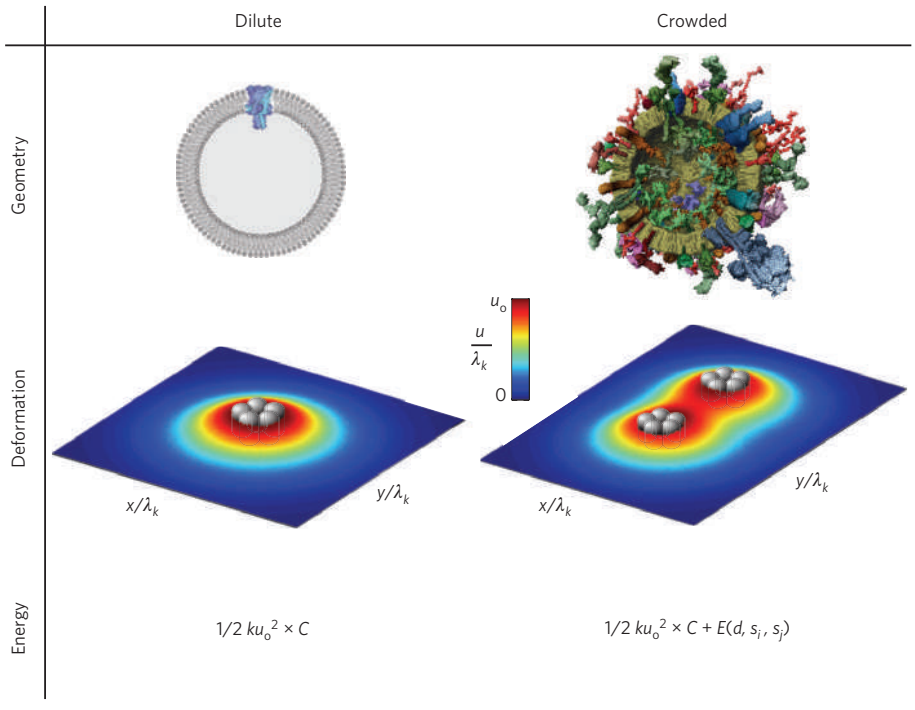


Figure 2 | Geometry, deformations and energies of dilute and crowded membranes. The two columns correspond to the dilute (proteins do not interact) and crowded (proteins interact) membrane limits. Each column shows the class of geometries found, a diagram of the deformation field in the vicinity of the proteins, and a mathematical description of the energies. For the isolated channel in the dilute limit, the deformation height, u, surrounding a given membrane protein has an elastic decay length, λ_k, that is smaller than the protein size. The deformation energy around a protein depends on a generic 'spring constant', k, determined from membrane properties. The deformation energy scales quadratically with hydrophobic mismatch, u_o, and scales approximately linearly with protein circumference, C. For crowded membranes, proteins have a sufficiently small separation distance ($d \approx \lambda_k$) that the annulus of deformed material around the proteins overlaps, resulting in an interaction energy that depends on the conformational state, s_i, of the ith protein.

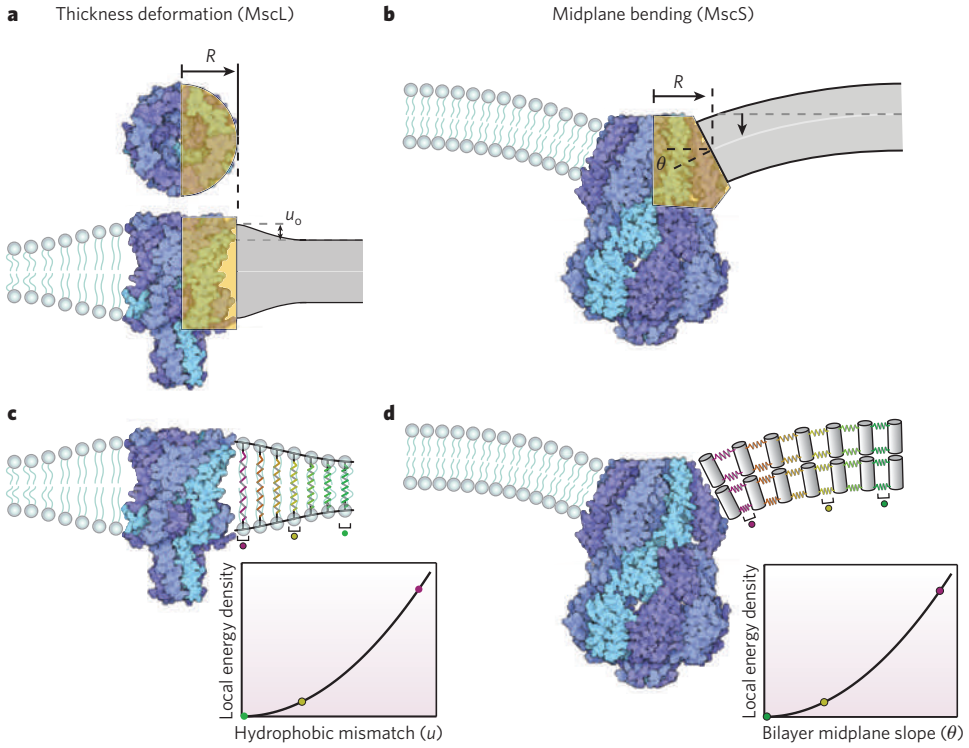


Figure 3 | Structure and energy at the protein–lipid interface. **a**, Atomic-level structure[63] and an elastic idealization of the mechanosensitive channel of large conductance (MscL) as a rigid cylinder with hydrophobic mismatch at the protein–lipid interface. R, effective radius of channel used in elastic model. u_o, hydrophobic mismatch between protein and equilibrium bilayer thickness. **b**, Atomic-level structure[64] and an elastic idealization of the mechanosensitive channel of small conductance (MscS) as a wedge with a slope that glues continuously onto the surrounding lipids. θ, midplane bending angle at protein–lipid interface. **c**, Membrane distortion and corresponding free energy of deformation per unit area of membrane surrounding MscL. **d**, Membrane distortion and corresponding free energy of deformation per unit area of membrane surrounding MscS. In **c** and **d** the elastic response of the lipids is captured with springs, and the colour coding indicates the local strain energy density at different distances from the proteins.

the gramicidin dimer[14]. Furthermore, using gramicidin A enantiomers as sensors for membrane mechanical properties, the small molecule capsaicin has been shown to target and trigger the pain receptor TRPV1 indirectly by decreasing the bending modulus of lipid bilayers in a concentration-dependent manner (not with a certain fixed stoichiometric relation between toxins and each channel, but progressively by altering the membrane's mechanical response)[11]. Conversely, voltage-dependent sodium channels are inactivated by capsaicin with no significant change to the conductance properties of the channels, but by an alteration of the gating voltage itself, suggesting that even channels that are not mechanically gated may still be subject to the effects of membrane mechanics through alterations of membrane properties[15–17]. In addition, it seems that some peptide toxins target multiple types of stretch-activated cation channels, not by changing membrane properties per se but by changing the effective boundary conditions at or near the protein–lipid interface[12]. This is yet another generic method by which membrane mechanics can couple to protein function (Fig. 1b). In particular, it seems that either enantiomer of a peptide toxin is localized in the membrane close to the channel and shifts its dose–response curve.

The experiments described above suggest ways of using quantitative models to explore the connection between membrane-protein function and the mechanics of the surrounding membrane. A useful starting point to flesh out a quantitative picture of such membranes is provided by simple order-of-magnitude estimates, and the derivation of scaling laws, for the free-energy costs associated with membrane deformations. For example, a simple census gives a sense of how many lipids surround each membrane protein, how far apart those proteins are in the membrane, and what this might imply about membrane-mediated interactions and corresponding cooperativity in protein function.

Experiments on the occupancy of biological membranes by lipids and their protein partners provide a useful place to start[18]. As shown in Fig. 2, proteomic and lipidomic approaches have made it possible to survey the protein and lipid content of biological membranes. In the case shown in Fig. 2, a survey of the contents of a synaptic vesicle reveals a crowded and heterogeneous medium. Indeed, as noted in the presentation of the original experiments: "A picture is emerging in which the membrane resembles a cobblestone pavement, with the proteins organized in patches that are surrounded by lipidic rims, rather than icebergs floating in a sea of lipids"[18].

The synaptic vesicle in Fig. 2 tells a similar story to results from other biological membranes, such as bacterial membranes or the protein census of the red-blood-cell membrane[19,20]. The essence of the various membrane inventories is that biological membranes are as much protein as they are lipid, with typical protein:lipid mass ratios of around 60:40 (refs 19, 20). There are many ways to estimate the mean spacing between membrane proteins, and we can quibble over the details, but the message is always the same: biological membranes are crowded. The mean centre-to-centre spacing between proteins is estimated at about 10 nm (comparable to the distance between proteins in the cytoplasm[21,22]), which tells us that these proteins might be able to influence each other through the intervening membrane.

A variety of theoretical tools can be used to explore the interactions of proteins and the surrounding membrane. Two of the most important classes of analysis of the link between structure and function are atomistic models, in which every atom is treated explicitly, and continuum elasticity models, in which the molecules of interest are represented by field variables that describe the height and thickness of the bilayer at each point. Although both are important, estimates can be built using simple arguments from elasticity. The conclusions are largely indifferent to the details of how the energetics of the composite lipid and membrane protein system are treated, and an atomistic analysis would yield the same general picture of a deformed footprint of material around the protein of interest, as indicated in Fig. 2. Even so, atomistic analyses can reveal features of membrane-protein function that are inaccessible to continuum analysis; several representative examples can be found in refs 23–26.

Additionally, certain theoretical constructs offer a correspondence between atomistic and continuum analysis. For instance, lipid pressure profiles are the statistical representation of fully atomistic bilayer forces, where the integral moments of the pressure profiles yield the continuum properties of lateral tension, bending rigidity, and monolayer and bilayer spontaneous curvature[27–29]. We refer to the generality of elasticity because the key ideas have to do with the kinds of generic, geometric perturbations on the lipids that can result from the presence of a membrane protein and the energetic consequences of the perturbations, especially where the membrane protein undergoes a conformational change in the course of its functional activity. The key ideas are indicated in Fig. 2, where both the 'dilute' and 'crowded' limits

Box 1 | Constants and scales involved in protein–lipid interactions

This Box provides typical values of some key mechanical properties relating to the bilayer and illustrates how the protein–lipid boundary conditions are altered by conformational changes. It also gives approximate analytical expressions and numerical estimates for the energetic costs of different types of bilayer deformation[38,40,42].

Bilayer properties

Here are some typical values of the bending and stretch stiffness, as well as the membrane thickness, of phospholipid bilayers. These lead to a relatively constant value for the elastic decay length for thickness deformations. Lateral tension and spontaneous curvature depend heavily on the osmotic conditions and amphiphilic composition of the bilayer, respectively, and so are quoted over a range. Variability in lateral tension also leads to a range of values for the midplane decay length.

Bending stiffness, $\kappa = 20\ k_B T$,
where k_B is the Boltzmann constant (1.38×10^{-23} J/K)
Stretch stiffness, $k_s = 60\ k_B T\ \mathrm{nm}^{-2}$
Thickness, $h_0 = 4$ nm
Tension, $\tau = 10^{-4}$–$1\ k_B T\ \mathrm{nm}^{-2}$
Spontaneous curvature, $C_0 = 0$–$0.04\ \mathrm{nm}^{-1}$
Transmembrane potential, $V_m = -40$ mV
Mismatch decay length, $\lambda_k = (\kappa h_0^2/4k_s)^{1/4} = 1.1$ nm
Midplane decay length, $\lambda_\tau = (\kappa/\tau)^{1/2} = 5$–$500$ nm

Channel properties

These are estimates of the change in radius, area and hydrophobic mismatch for the two-state conformational change of the mechanosensitive channel of large conductance, MscL. A hypothetical value for the change in midplane tilt at the protein–lipid interface is given for a generic two-state protein. The gating charge of a typical voltage-gated potassium channel is given for use in an energetic comparison.

Radius (closed–open), $R = 2.5$–3.5 nm
Area variation, $\Delta A \approx 20\ \mathrm{nm}^2$
Thickness variation, $|\Delta u_0| = 0.8$ nm
Midplane tilt variation, $|\Delta\theta| = \pi/4$
Charge variation, $\Delta Q = 12e$,
where e is the elementary unit of charge (1.6×10^{-19} C)

Conformational energy

We have used approximate analytical expressions for bilayer mechanical properties and interfacial boundary conditions to estimate the energetic costs of bilayer thickness deformation, protein area change, midplane deformations, bilayer spontaneous curvature coupling to the midplane and, for comparison, the free energy of voltage gating. These expressions are given in the limit $\lambda_k < R < \lambda_\tau$.

Thickness variation, $\Delta E_u = 2^{1/2}\pi\kappa(R/\lambda_k)(\Delta u_0/\lambda_k)^2 = 15\ k_B T$
Area variation, $\Delta E_\tau = \tau\Delta A = 10^{-3}$–$20\ k_B T$
Tilt variation where $C_0 = 0$, $\Delta E_\theta = \pi\tau(R\theta)^2 \ln(\lambda_\tau/R) = 10^{-3}$–$7\ k_B T$
Tilt variation where $C_0 \neq 0$, $\Delta E_\theta = 2\pi\kappa C_0 R\theta = 10\ k_B T$
Charge variation, $\Delta E_v = \Delta Q V_m = 20\ k_B T$

show how membrane proteins perturb the surrounding lipids (and each other, if the membrane is sufficiently crowded).

Elasticity and the isolated channel

To understand the interplay between ion channels and the surrounding lipids, consider an idealized isolated channel, like that in the top left-hand corner of Fig. 2, in a single-component lipid bilayer. Such simplifications fall short of the rich and varied landscape inhabited by channels in real cell membranes, but they can still provide useful mechanistic insights into how membrane proteins function when the protein 'footprint' changes with the conformation.

Figure 2 can help us predict the results of a mathematical description of these channels using elasticity theory. For example, an ion channel might change its external radius during gating or the thickness of its hydrophobic region[4]. The key point is that the region of the membrane coloured red in Fig. 2 corresponds to membrane that is deformed (not in the relaxed state it would adopt if no protein were present). This region of deformed material costs a certain amount of deformation free energy. Furthermore, when the protein changes conformation, the annulus of deformed material changes, and so does the free-energy penalty.

Membrane as an elastic sheet

A convenient model for describing the interaction between membrane proteins and the surrounding membrane is to consider the membrane as a continuous elastic medium[30–32]. The idea behind this description is that there is a cost in terms of free energy that must be paid for perturbing the lipid bilayer away from some undeformed reference state, as indicated by the coloured springs in Fig. 3c, d. We emphasize several key modes of deformation (notably hydrophobic mismatch and midplane bending) and their corresponding free-energy cost. There is a well-defined mathematical theory of the free energy of membrane deformation[33–36], but here we will emphasize instead a qualitative and intuitive description of these theoretical results.

As highlighted in Fig. 3, different types of membrane deformation have a free-energy cost that can be calculated in the form of energy density (the free energy per unit area of deformed membrane). For example, in the case of a hydrophobic mismatch, where there is a free-energy penalty associated with 'gluing' the hydrophobic lipid tails to the hydrophobic region of the membrane protein, the free-energy density increases as the square of the hydrophobic mismatch[36–38]. Similarly, some membrane proteins will bend the membrane bilayer in their vicinity, incurring another class of free-energy cost[37,38]. This idea is represented in Fig. 3 by thinking of the membrane as a set of generalized springs. For every patch of area on the membrane, we can ask how different the thickness is from the equilibrium thickness, and how much the membrane is bent away from the flat state (in which there is no spontaneous curvature). Given the answer to these geometric questions, we can use these generalizations of Hooke's law to assign an energy density (the energy per unit area) to each patch of membrane, so we can find the total free-energy cost by summing over all such patches.

Energy and length scales

Essential to gauging the importance of the interplay between lipids and membrane proteins, and of any subsequent membrane-mediated interactions, are estimates of the energetic costs of membrane deformation and the size of the region over which that deformation occurs. Membranes are composed of a plethora of different lipid species, but on the length and time scales of interest, several coarse-grained continuum material properties emerge[33]. For a homogenous lipid phase, these material parameters are the bending stiffness (with units of energy, measured here in terms of $k_B T$, where k_B is the Boltzmann constant and $T = 300$ K), the stretch stiffness and the membrane tension (both with units of energy per unit area, measured here in units of $k_B T\ \mathrm{nm}^{-2}$), the bilayer thickness (measured in nm) and the spontaneous curvature of the membrane (measured in nm^{-1}) (Box 1). We will proceed as though these parameters were true material constants, although the situation is more subtle because the lipid environment surrounding a protein can change with its conformational state, and so too can these parameters.

Within the continuum elastic equations that describe the membrane deformation, certain 'natural' length and energy scales emerge that can serve as a guide to our thinking and provide intuition about the relative importance of different effects[30,39,40] (Box 1). Midplane bending and thickness deformation share a common energy scale proportional to the bending stiffness. If all other membrane and protein properties are fixed, stiffer membranes will cost more energy to deform. Likewise, both modes of deformation share a common energy scaling with changes in the relevant boundary condition at the protein–lipid interface[41–43]. In midplane deformation, protein 'shape' dictates the angle at which

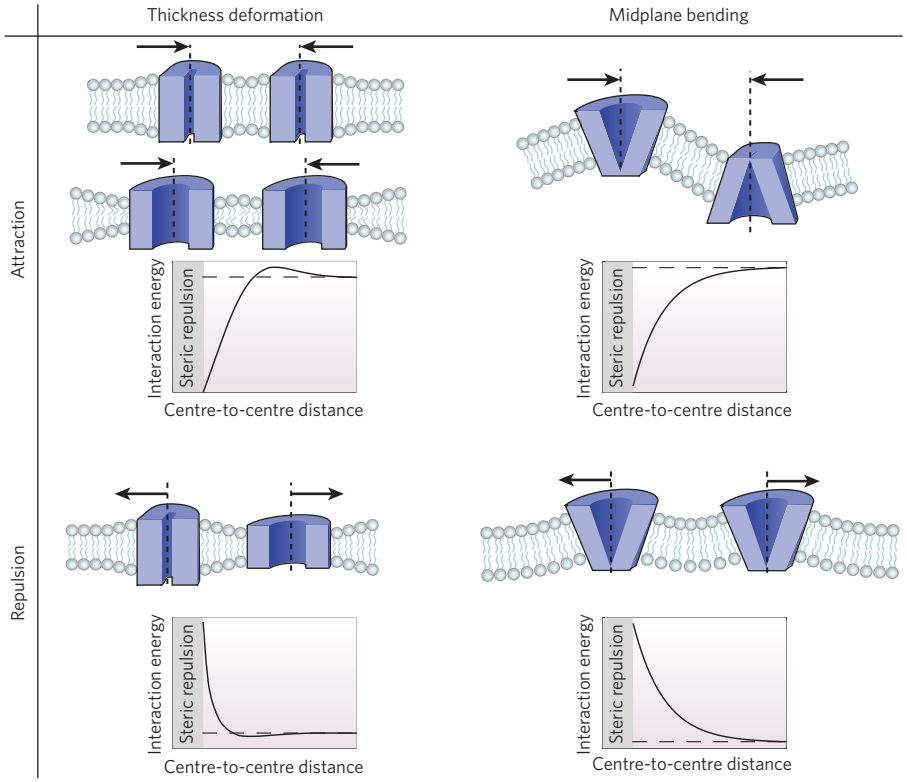


Figure 4 | Membrane–protein interactions and conformational state. Overlap in the deformed membrane between proteins can cause attraction or repulsion over distances comparable to the elastic decay length. The interaction energy between membrane proteins depends on their conformational state and can induce cooperative conformational changes.

the membrane contacts the protein (Fig. 3); the deformation energy here increases quadratically in the contact angle, so it acts rather like a classical Hookean spring. Similarly, in thickness deformation, as the degree of hydrophobic mismatch between the embedded protein and the bilayer increases, the deformation energy increases quadratically. In practice, a reasonable estimate of deformation energy around a protein is 10 k_BT in either case[40,43] (Box 1).

Membrane elasticity and mechanosensitive channel gating

The ideas developed above can be used to understand the origin of the effects shown in Fig. 1. To see how, we use ideas from statistical mechanics to write the open probability of a channel as a function of the driving force of interest[31,43,44], resulting in:

$$p_{\mathrm{open}} = \frac{e^{-\beta\varepsilon_{\mathrm{open}}}}{e^{-\beta\varepsilon_{\mathrm{closed}}} + e^{-\beta\varepsilon_{\mathrm{open}}}} \qquad (1)$$

where β is $1/k_BT$ and $\varepsilon_{\mathrm{open}}$ and $\varepsilon_{\mathrm{closed}}$ refer to the free energies of the open and closed states, respectively. The free energies of the open and closed states are tuned by changing the contribution of the driving force to these two energies. This result can be specialized to the case of tension-driven ion-channel gating by noting that the energies of the open and closed states are dictated by the coupling to the tension and by the free-energy cost of the annulus of deformed material surrounding the channel, such that

$$p_{\mathrm{open}} = \frac{1}{1 + e^{-\beta(\tau\Delta A + \Delta\varepsilon_{\mathrm{membrane}} + \Delta\varepsilon_{\mathrm{protein}})}} \qquad (2)$$

This kind of analysis can respond to experiments like those in Fig. 1a. The term $\tau\Delta A$ corresponds to the driving force that favours the open state (τ is the membrane tension and ΔA is the change in protein area). However, this driving force must compete with the free-energy penalty associated with the membrane deformation footprint ($\Delta\varepsilon_{\mathrm{membrane}}$) introduced in Fig. 2. We also include the energy difference ($\Delta\varepsilon_{\mathrm{protein}}$) between the open and closed states associated with the protein's internal degrees of freedom. Direct comparison with the experimental results in Fig. 1a is difficult because only the pipette pressure is reported experimentally, whereas the membrane tension is the key driving force[8]. Experiments that measure bilayer tension show that gating the mechanosensitive channel of large conductance (MscL) requires ~2 $k_BT\,\mathrm{nm}^{-2}$, depending on the lipid, and is accompanied by a change in area between open and closed conformations, ΔA, of around 20 nm^2, corresponding to a gating energy, $\tau\Delta A = 40\ k_BT$. As expected, this energy is much larger than the thermal energy, so spontaneous channel opening under low tension rarely occurs. Box 1 provides estimates for the membrane-associated energy penalty, and quantitatively shows the main point of this Review, namely that this membrane-deformation free energy can be reasonably expected to compete with the driving force, and can thereby influence the protein conformation.

As illustrated above, there are generic reasons to expect that for any membrane protein that alters its deformation footprint during a conformational change, protein function will be dependent on the structure of the membrane (and possibly on the tension as well). Our discussion has focused on general principles, rather than specific examples, and we have not dwelled on the contribution to the free energy from protein conformation ($\Delta\varepsilon_{\mathrm{protein}}$ in eqn (2)). It should be emphasized that although lipid membranes generally exert a (composition-dependent) mechanical stress on embedded membrane proteins, the way that a given protein responds to this stress is highly specific. The mechanical stress from membrane deformation is likely to have little or no effect on proteins that offer a rigid (non-deformable) interface to the lipid membrane, or on proteins that show high affinity for particular lipids and that will not be influenced by the overall membrane composition if those lipids are present in the membrane.

Interacting membrane proteins and cooperativity

Several key insights from the discussion of the isolated channel can help us examine what happens when there are multiple membrane proteins. When different membrane proteins are within several elastic decay lengths of each other, they will interact. However, depending on protein shape, these membrane-mediated interactions may result in either attraction or repulsion[30,39,45–48]. Furthermore, at these small length

scales, where thermal fluctuations are important, membrane-mediated interactions between proteins can arise from rigid proteins perturbing the allowed jiggling motions of the membrane (the Casimir effect)[49,50]. These interactions are potentially long-ranged, but might be small, and their physiological relevance to membrane proteins has not been demonstrated.

The rules introduced in Box 1 hold during membrane-mediated protein interactions, except that we must compare the spatial extent of the deformation field with the distance between proteins. Each type of deformation has a length scale over which the membrane returns to its unperturbed state[30,39,40], although the interactions from thickness and midplane deformations behave qualitatively differently. The length scale of thickness deformations is mainly constant[33] and short (about a nanometre), but the length scale of midplane bending interactions is variable and longer (5–500 nm)[50], and tends to be weaker. These characteristic length scales are important in determining the balance between membrane deformation and the generic driving force that determines the conformation of the interacting proteins.

Biological membranes at physiological temperatures are generally in a fluid state, so both lipids and proteins can diffuse laterally (provided that they do not strongly interact with the cytoskeleton[51]). The diffusing proteins can be thought of as a two-dimensional gas with an entropic tension, equivalent to the pressure in a gas, that acts on the external surface of each protein in the membrane because the remaining proteins are jiggling around in the area available to them. The opening of membrane channels is typically associated with a change in channel area. One potential consequence of membrane crowding is that this conformational change could cause a change in free energy that is associated with the area available to the rest of the proteins, resulting in depletion forces.

Other effects can also arise through explicit interactions between adjacent channels, as shown in Fig. 4. The idea that the conformational states of two similar proteins can be coupled by the bilayer follows naturally from the discussion of membrane deformation and has been explored in detail elsewhere. Two proteins in proximity (within a few elastic decay lengths) will have regions of bilayer deformation that overlap, so one protein indirectly affects another through the lipids that surround them both[30,39,45–48,52,53]. One interesting outcome of these interactions is cooperative channel gating, as a conformational change in one protein will be 'felt' energetically through the surrounding lipids, influencing another protein's preference for a particular conformation[54]. Alternatively, the binding of membrane-associated proteins may impose boundary conditions that deform the membrane midplane, similar to the embedded proteins mentioned previously[55]. Several studies have shown that large-scale membrane deformations, such as budding[56] and tubulation[57], result from the collective mechanical interactions of such proteins[48,58–60], as reviewed in ref. 61. Here, too, certain rules emerge that depend on the nature of the interaction.

Within these elastic models, proteins that cause thickness deformation tend to attract each other if they both increase or both decrease the bilayer thickness. Conversely, if one protein thickens the bilayer and another thins it, they will repel each other. Proteins that bend the midplane of the bilayer have the opposite behaviour: those that bend the bilayer in the same direction tend to repel each other, whereas those that bend the bilayer in opposite directions tend to attract each other[52]. In either case, attraction arises because the amount of deformed material between the proteins decreases when proteins are in close proximity, lowering the deformation free energy. Proteins that attract each other have more deformation overlap and are more likely to be found within each other's circle of influence, so they have more strongly coupled conformations.

Much work has been done on the nature of these interactions as an organizing principle for lipids and proteins. Our emphasis here is on a second consequence of such interactions: their ability to induce cooperativity in the conformational changes of neighbouring membrane proteins. We will not discuss the details here, but the outcome of the interactions is the principle that if one channel decides to gate, this increases the likelihood that its neighbours will gate as well[54]. The more severe the bilayer deformation, the stronger the interaction will be between similar proteins, and the more tightly their conformations will couple.

Concluding perspective

A range of evidence for several different membrane proteins reveals the role played by the character of the surrounding membrane. We suggest that the natural regulatory effect of lipids on membrane-protein function can be used to dissect the structure–function relationship of membrane proteins. To make further progress in understanding the richness of the interface between membrane proteins and lipids, one useful avenue might be to exploit the generic predictions resulting from the kinds of theoretical analysis described here for the way in which conformational changes depend on the properties of the surrounding lipids and proteins. Beyond this, the role of membrane crowding should be explored more systematically because the proximity of membrane proteins may result in the same kinds of surprises already seen for crowding effects in the bulk setting[21,22,62]. ■

1. Engelman, D. M. Membranes are more mosaic than fluid. *Nature* **438,** 578–580 (2005).
2. Sukharev, S. I. *et al.* Energetic and spatial parameters for gating of the bacterial large conductance mechanosensitive channel, MscL. *J. Gen. Physiol.* **113,** 525–540 (1999).
3. Perozo, E. *et al.* Physical principles underlying the transduction of bilayer deformation forces during mechanosensitive channel gating. *Nature Struct. Biol.* **9,** 696–703 (2002).
This paper describes experiments that show how the gating behaviour of mechanosensitive channels depends on the acyl chain lengths of the lipids around them.
4. Andersen, O.S. & Koeppe, R.E. II Bilayer thickness and membrane protein function: an energetic perspective. *Annu. Rev. Biophys. Biomol. Struct.* **36,**107–130 (2007).
This useful review provides a variety of examples of the connection between the lipid environment and the function of membrane-bound proteins.
5. Marsh, D. Protein modulation of lipids, and vice-versa, in membranes. *Biochim. Biophys. Acta* **1778,** 1545–1575 (2008).
6. Killian, J. A. Hydrophobic mismatch between proteins and lipids in membranes. *Biochim. Biophys. Acta* **1376,** 401–415 (1998).
7. Petsko, G. & Ringe, D. *Protein Structure and Function* (New Science Press, 2004).
8. Moe, P. & Blount, P. Assessment of potential stimuli for mechano-dependent gating of MscL: effects of pressure, tension, and lipid headgroups. *Biochemistry* **44,** 12239–12244 (2005).
9. Lee, A. G. How lipids affect the activities of integral membrane proteins. *Biochim. Biophys. Acta* **1666,** 62–87 (2004).
This review discusses how specific residues in the structure of transmembrane proteins determine hydrophobic thickness and addresses the topic of site-specific lipid binding.
10. Valiyaveetil, F. I., Zhou, Y. & MacKinnon, R. Lipids in the structure, folding, and function of the KcsA K^+ channel. *Biochemistry* **41,** 10771–10777 (2002).
11. Lundbaek, J. A. *et al.* Capsaicin regulates voltage-dependent sodium channels by altering lipid bilayer elasticity. *Mol. Pharmacol.* **68,** 680–689 (2005).
12. Suchyna, T. M. *et al.* Bilayer-dependent inhibition of mechanosensitive channels by neuroactive peptide enantiomers. *Nature* **430,** 235–240 (2004).
13. Jensen, M. O. & Mouritsen, O. G. Lipids do influence protein function — the hydrophobic matching hypothesis revisited. *Biochim. Biophys. Acta* **1666,** 205–226 (2004).
This review discusses the qualitative effects of thickness mismatch on membrane proteins and summarizes evidence of the effects of bilayer thickness on enzymatic activity.
14. Goulian, M. *et al.* Gramicidin channel kinetics under tension. *Biophys. J.* **74,** 328–337 (1998).
15. Schmidt, D. & MacKinnon, R. Voltage-dependent K^+ channel gating and voltage sensor toxin sensitivity depend on the mechanical state of the lipid membrane. *Proc. Natl Acad. Sci. USA* **105,** 19276–19281 (2008).
16. Calabrese, B. *et al.* Mechanosensitivity of N-type calcium channel currents. *Biophys. J.* **83,** 2560–2574 (2002).
17. Morris, C. E. & Juranka, P. F. Nav channel mechanosensitivity: activation and inactivation accelerate reversibly with stretch. *Biophys. J.* **93,** 822–833 (2007).
18. Takamori, S. *et al.* Molecular anatomy of a trafficking organelle. *Cell* **127,** 831–846 (2006).
This paper provides an experimental treatment of the crowded nature of biological membranes.
19. Dupuy, A. D. & Engelman, D. M. Protein area occupancy at the center of the red blood cell membrane. *Proc. Natl Acad. Sci. USA* **105,** 2848–2852 (2008).
20. Mitra, K. *et al.* Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *Proc. Natl Acad. Sci. USA* **101,** 4083–4088 (2004).
This experimental *tour de force* shows that protein content modulates bilayer thickness across different organellar membranes by the hydrophobic mismatch principle.
21. Zimmerman, S. B. & Minton, A. P. Macromolecular crowding: biochemical, biophysical, and physiological consequences. *Annu. Rev. Biophys. Biomol. Struct.* **22,** 27–65 (1993).
22. Ellis, R. J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* **26,** 597–604 (2001).
23. Bond, P. J. & Sansom, M. S. Bilayer deformation by the Kv channel voltage sensor domain revealed by self-assembly simulations. *Proc. Natl Acad. Sci. USA* **104,** 2631–2636 (2007).
24. Chen, X. *et al.* Gating mechanisms of mechanosensitive channels of large conductance, I: a continuum mechanics-based hierarchical framework. *Biophys. J.* **95,** 563–580 (2008).
25. Elmore, D. E. & Dougherty, D. A. Investigating lipid composition effects on the mechanosensitive channel of large conductance (MscL) using molecular dynamics

simulations. *Biophys. J.* **85,** 1512–1524 (2003).
26. Jeon, J. & Voth, G. A. Gating of the mechanosensitive channel protein MscL: the interplay of membrane and protein. *Biophys. J.* **94,** 3497–3511 (2008).
27. Marsh, D. Lateral pressure profile, spontaneous curvature frustration, and the incorporation and conformation of proteins in membranes. *Biophys. J.* **93,** 3884–3899 (2007).
28. Cantor, R. The influence of membrane lateral pressures on simple geometric models of protein conformational equilibria. *Chem. Phys. Lipids* **101,** 45–56 (1999).
29. Marsh, D. Elastic curvature constants of lipid monolayers and bilayers. *Chem. Phys. Lipids* **144,** 146–159 (2006).
This review compiles measurements of membrane elastic properties across different lipid types and builds scaling laws that connect different modes of bilayer deformation to bulk bilayer properties.
30. Dan, N., Pincus, P. & Safran, S. Membrane-induced interactions between inclusions. *Langmuir* **9,** 2768–2771 (1993).
31. Markin, V. S. & Sachs, F. Thermodynamics of mechanosensitivity. *Phys. Biol.* **1,** 110–124 (2004).
32. Mouritsen, O. G. & Bloom, M. Models of lipid-protein interactions in membranes. *Annu. Rev. Biophys. Biomol. Struct.* **22,** 145–171 (1993).
33. Boal, D. *Mechanics of the Cell* (Cambridge University Press, 2002).
34. Gruner, S. M. Intrinsic curvature hypothesis for biomembrane lipid composition: a role for nonbilayer lipids. *Proc. Natl Acad. Sci. USA* **82,** 3665–3669 (1985).
35. Mouritsen, O. G. & Bloom, M. Mattress model of lipid–protein interactions in membranes. *Biophys. J.* **46,** 141–153 (1984).
36. Huang, H. W., Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.* **50,** 1061–1070 (1986).
37. Turner, M. S. & Sens, P. Gating-by-tilt of mechanically sensitive membrane channels. *Phys. Rev. Lett.* **93,** 118103 (2004).
38. Wiggins, P. & Phillips, R. Membrane-protein interactions in mechanosensitive channels. *Biophys. J.* **88,** 880–902 (2005).
39. Turner, M. S. & Sens, P. Inclusions on fluid membranes anchored to elastic media. *Biophys. J.* **76,** 564–572 (1999).
40. Wiggins, P. & Phillips, R. Analytic models for mechanotransduction: gating a mechanosensitive channel. *Proc. Natl Acad. Sci. USA* **101,** 4071–4076 (2004).
41. Dan, N. & Safran, S. A. Effect of lipid characteristics on the structure of transmembrane proteins. *Biophys. J.* **75,** 1410–1414 (1998).
42. Nielsen, C., Goulian, M. & Andersen, O. S. Energetics of inclusion-induced bilayer deformations. *Biophys. J.* **74,** 1966–1983 (1998).
43. Ursell, T., Reeves, D., Wiggins, P. & Phillips, R. in *Mechanosensitive Ion Channels* (ed. Kamkin, A. & Kisileva, I.) 37–70 (Springer, 2008).
44. Cantor, R. S. Lateral pressures in cell membranes: A mechanism for modulation of protein function. *J. Phys. Chem. B* **101,** 1723–1725 (1997).
45. Bruinsma, R., Goulian, M. & Pincus, P. Self-assembly of membrane junctions. *Biophys. J.* **67,** 746–750 (1994).
46. Gil, T. *et al.* Theoretical analysis of protein organization in lipid membranes. *Biochim. Biophys. Acta* **1376,** 245–266 (1998).
47. Muller, M. M., Deserno, M. & Guven, J. Interface-mediated interactions between particles: a geometrical approach. *Phys. Rev. E* **72,** 061407 (2005).
48. Reynwar, B. J. et al. Aggregation and vesiculation of membrane proteins by curvature-mediated interactions. *Nature* **447,** 461–464 (2007).
49. Golestanian, R., Goulian, M. & Kardar, M. Fluctuation-induced interactions between rods on a membrane. *Phys. Rev. E* **54,** 6725–6734 (1996).
50. Goulian, M., Bruinsma, R. & Pincus, P. Long-range forces in heterogeneous fluid membranes. *Europhys. Lett.* **22,** 145–150 (1993).
51. Sheetz, M.P., Sable, J.E. & Dobereiner, H. G. Continuous membrane-cytoskeleton adhesion requires continuous accommodation to lipid and cytoskeleton dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 417–434 (2006).
52. Weikl, T. R., Kozlov, M. M. & Helfrich, W. Interaction of conical membrane inclusions: Effect of lateral tension. *Phys. Rev. E* **57,** 6988–6995 (1998).
53. Goforth, R. L. *et al.* Hydrophobic coupling of lipid bilayer energetics to channel function. *J. Gen. Physiol.* **121,** 477–493 (2003).
54. Sens, P., Johannes, L. & Bassereau, P. Biophysical approaches to protein-induced membrane deformations in trafficking. *Curr. Opin. Cell Biol.* **20,** 476–482 (2008).
55. Blood, P. D. & Voth, G. A. Direct observation of Bin/amphiphysin/Rvs (BAR) domain-induced membrane curvature by means of molecular dynamics simulations. *Proc. Natl Acad. Sci. USA* **103,** 15068–15072 (2006).
56. Reynwar, B. J. *et al.* Aggregation and vesiculation of membrane proteins by curvature-mediated interactions. *Nature* **447,** 461–464 (2007).
57. Arkhipov, A., Yin, Y. & Schulten, K. Four-scale description of membrane sculpting by BAR domains. *Biophys. J.* **95,** 2806–2821 (2008).
58. Kim, K. S., Neu, J. & Oster, G. Curvature-mediated interactions between membrane proteins. *Biophys. J.* **75,** 2274–2291 (1998).
59. Leibler, S. Curvature instability in membranes. *J. Physique* **47,** 507–516 (1986).
60. Sens, P. & Turner, M. S. Theoretical model for the formation of caveolae and similar membrane invaginations. *Biophys. J.* **86,** 2049–2057 (2004).
61. Ursell, T. *et al.* Cooperative gating and spatial organization of membrane proteins through elastic interactions. *PLoS Comput. Biol.* **3,** e81 (2007).
62. Zhou, H. X., Rivas, G. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **37,** 375–397 (2008).
63. Chang, G. *et al.* Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282,** 2220–2226 (1998).
64. Bass, R. B. *et al.* Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. *Science* **298,** 1582–1587 (2002).

Acknowledgements We are grateful to many people for insights and comments. In particular, we thank O. Andersen, M. Bialecka, F. Brown, N. Dan, L. Haswell, S. Johnson, J. Kondev, R. MacKinnon, C. Morris, D. Rees, D. Reeves, F. Sachs, D. Schmidt, S. Scheuring, S. Sukharev, M. Turner, S. White and M. Widom. We are also grateful to several anonymous reviewers for insightful comments. In addition, we are grateful to the US National Science Foundation and the National Institutes of Health for support through NIH Award number R01 GM084211 and the Director's Pioneer Award. The reference list for this article was constrained by length limits, and as a result the references cited here are representative rather than comprehensive. We apologize to those whose references are not cited as a result of either space limitations or our ignorance.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence should be addressed to R.P. (phillips@pboc.caltech.edu).

ARTICLES

Bmi1 regulates mitochondrial function and the DNA damage response pathway

Jie Liu[1]*, Liu Cao[1]*, Jichun Chen[2], Shiwei Song[1], In Hye Lee[1], Celia Quijano[1], Hongjun Liu[1], Keyvan Keyvanfar[2], Haoqian Chen[1], Long-Yue Cao[1], Bong-Hyun Ahn[1], Neil G. Kumar[1,3], Ilsa I. Rovira[1], Xiao-Ling Xu[4], Maarten van Lohuizen[5], Noboru Motoyama[6], Chu-Xia Deng[4] & Toren Finkel[1]

Mice deficient in the Polycomb repressor Bmi1 develop numerous abnormalities including a severe defect in stem cell self-renewal, alterations in thymocyte maturation and a shortened lifespan. Previous work has implicated de-repression of the *Ink4a/Arf* (also known as *Cdkn2a*) locus as mediating many of the aspects of the *Bmi1*$^{-/-}$ phenotype. Here we demonstrate that cells derived from *Bmi1*$^{-/-}$ mice also have impaired mitochondrial function, a marked increase in the intracellular levels of reactive oxygen species and subsequent engagement of the DNA damage response pathway. Furthermore, many of the deficiencies normally observed in *Bmi1*$^{-/-}$ mice improve after either pharmacological treatment with the antioxidant *N*-acetylcysteine or genetic disruption of the DNA damage response pathway by Chk2 (also known as Chek2) deletion. These results demonstrate that Bmi1 has an unexpected role in maintaining mitochondrial function and redox homeostasis and indicate that the Polycomb family of proteins can coordinately regulate cellular metabolism with stem and progenitor cell function.

Bmi1 is a member of the Polycomb family of transcriptional repressors that mediate gene silencing by regulating chromatin structure and is essential for the maintenance and self-renewal of both haematopoietic and neural stem cells[1,2]. Mice deficient in Bmi1 have a number of defects including severe neurological abnormalities, alterations in various haematopoietic cell lineages, a generalized failure-to-thrive and a markedly shortened lifespan. Previous studies have documented that approximately 50% of the knockout mice die before the completion of weaning and the remaining 50% succumb anywhere from 3 to 20 weeks of age[3]. In an effort to understand how the effects of Bmi1 are mediated, significant attention has been placed on the repression of the *Ink4a/Arf* locus, which encodes two separate tumour suppressors and cell cycle regulators, $p16^{Ink4a}$ and $p19^{Arf}$ (refs 4–7). The importance of the transcriptional repression of the *Ink4a/Arf* locus is underscored by the observation that mice deficient in Bmi1 and also lacking $p16^{Ink4a}$, $p19^{Arf}$ or both of these gene products develop less severe neurological and haematological abnormalities than *Bmi1*$^{-/-}$ animals[4–6].

Although improved, mice lacking Bmi1 as well as Ink4a/Arf remain significantly smaller than wild-type littermates and their overall survival is similar to that of *Bmi1*$^{-/-}$ mice[6]. In addition, examination of the number of peripheral nucleated thymocytes or splenocytes, as well as the cellular composition of the bone marrow microenvironment itself, remains significantly altered in combined Bmi1/Ink4a/Arf-deficient mice[8]. These results therefore indicate that, besides repression of the *Ink4a/Arf* locus, additional Bmi1-regulated pathways undoubtedly exist. Here we demonstrate that Bmi1 can separately regulate mitochondrial function, reactive oxygen species (ROS) levels and the activation of the DNA damage response (DDR) pathway.

Bmi1 regulates ROS levels

Mice deficient in either the ataxia telangiectasia mutated (Atm) gene product or the FOXO family of transcription factors demonstrate a rapid postnatal decline in haematopoietic stem cell (HSC) number that in each case is associated with a rise in ROS levels within the c-KIT$^+$Sca-1$^+$Lin$^-$ (LSK) bone marrow cell population that includes HSCs[9,10]. We therefore sought to test whether a similar rise in ROS levels was evident in *Bmi1*$^{-/-}$ cells. Assessment of unfractionated bone marrow, purified LSK cells and bone marrow cells expressing SLAM family receptors previously shown to allow for enrichment of long-term HSC cells (LT-HSCs)[11] demonstrated that the absence of *Bmi1*$^{-/-}$ resulted in increased levels of ROS (Fig. 1a and Supplementary Fig. 1). A similar increase in ROS levels was also evident in various other cell types known to be impaired in *Bmi1*$^{-/-}$ mice, including freshly isolated thymocytes (Fig. 1b and Supplementary Fig. 1).

In an effort to explain this observed rise in ROS, we made use of several previous gene expression studies that have identified a multitude of Polycomb target genes[1,12,13]. Among these identified targets were numerous genes either involved directly in ROS generation or that localize to the mitochondria and are known to affect mitochondrial function. Given the scarcity of LT-HSCs in *Bmi1*$^{-/-}$ mice, we chose to analyse gene expression in *Bmi1*$^{-/-}$ thymocytes, because previous reports have demonstrated a marked perturbation of this cell population in Bmi1-deficient mice[3,4] and our data indicated that these cells had marked differences in ROS levels. As noted in Fig. 1c, *Bmi1*$^{-/-}$ thymoctyes de-repressed a number of previously identified Polycomb-regulated gene products that can regulate intracellular redox homeostasis. Given that mitochondria can produce oxidants and are in turn quite sensitive to their damaging effects[14], we sought to analyse the mitochondrial function of *Bmi1*$^{-/-}$ cells. Intact *Bmi1*$^{-/-}$ thymocytes had both reduced basal mitochondrial oxygen consumption and reduced mitochondrial oxidative capacity (Fig. 1d). Basal ATP levels were also significantly reduced in *Bmi1*$^{-/-}$ thymocytes (Fig. 1e) as well as in other tissues

[1]Translational Medicine Branch, National Heart Lung and Blood Institute, National Institutes of Health, [2]Hematology Branch, National Heart Lung and Blood Institute, National Institutes of Health, [3]Howard Hughes Medical Institute, NIH Research Scholar Program, [4]Genetics of Development and Disease Branch, National Institute of Diabetes and Digestive and Kidney Disease, National Institutes of Health, Bethesda, Maryland 20892, USA. [5]Division of Molecular Genetics, Netherlands Cancer Institute and Centre for Biomedical Genetics, 1066 CX Amsterdam, The Netherlands. [6]Department of Geriatric Medicine, National Institute for Longevity Sciences National Center for Geriatrics and Gerontology 36-3, Gengo, Morioka, Obu, Aichi 474-8522, Japan.
*These authors contributed equally to this work.

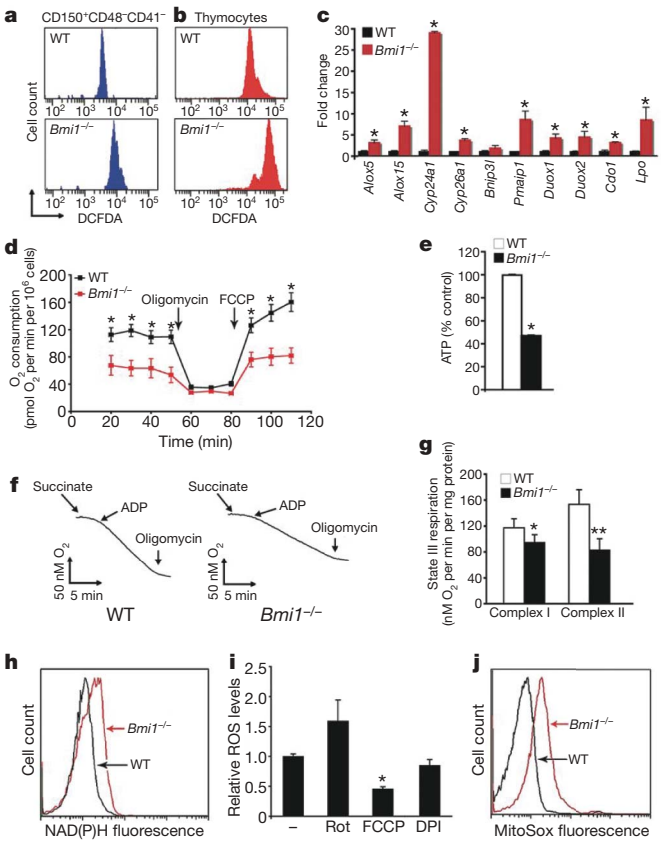


Figure 1 | The absence of Bmi1 increases ROS levels and alters mitochondrial function. **a**, Levels of ROS as assessed by DCFDA fluorescence in purified wild-type (WT) or *Bmi1*$^{-/-}$ bone marrow cells. **b**, ROS levels in freshly isolated thymocytes. **c**, Quantitative rtPCR expression analysis of gene products involved in redox homeostasis in either WT or *Bmi1*$^{-/-}$ thymocytes. Results are normalized to *Gapdh* expression ($n = 3$ animals per group, $^*P < 0.05$). **d**, Oxygen consumption in intact thymocytes under basal conditions, following the addition of the mitochondrial inhibitor oligomycin (0.5 μg ml^{-1}) or in the presence of the uncoupler FCCP (1 μM). Cells were obtained from $n = 4$ animals per genotype; $^*P < 0.02$. **e**, Relative ATP levels in freshly isolated thymocytes. Measurements were made in triplicate (mean and s.d.) with 4 animals per group; $^*P < 0.01$. **f**, Representative oxygen consumption from isolated heart mitochondria. **g**, Measured state III respiration for either complex-I- or complex-II-dependent respiration (mean and s.d.; $n = 4$ animals per group; $^*P < 0.05$, $^{**}P < 0.01$). **h**, NAD(P)H levels as assessed by endogenous fluorescence in WT and *Bmi1*$^{-/-}$ thymocytes. **i**, ROS levels in *Bmi1*$^{-/-}$ thymocytes in the presence or absence of the complex I inhibitor rotenone (Rot), the chemical uncoupler FCCP or the NADPH oxidase inhibitor DPI (mean and s.d.; $n = 3$; $^*P < 0.05$). **j**, Analysis of thymocytes using the redox fluorophore MitoSox Red.

(Supplementary Fig. 2). The observed differences in mitochondrial respiration and intracellular energetics were not however accompanied by any obvious differences in mitochondrial number, structure, biogenesis or rate of degradation (Supplementary Fig. 3).

Consistent with our intact cellular respiration measurements, we observed significant impairment in the function of purified *Bmi1*$^{-/-}$ mitochondria (Fig. 1f, g). The interruption of flow down the electron transport chain (ETC) evident in *Bmi1*$^{-/-}$ mitochondria is believed to be a major source for ROS generation. Impaired electron flow increases the likelihood of superoxide formation and is often accompanied by a build-up of mitochondrial reducing equivalents (for example, NADH). Consistent with such a mechanism, NAD(P)H levels were increased in *Bmi1*$^{-/-}$ thymocytes (Fig. 1h). To substantiate further the connection between Polycomb activity, mitochondrial function and oxidant stress, we next measured ROS levels in the setting of various inhibitors of mitochondrial function. As expected, using the complex I inhibitor rotenone to reduce electron flow increased ROS levels (Fig. 1i). In contrast, treatment with the pharmacological uncoupler FCCP, which specifically lowers mitochondrial membrane potential and hence augments ETC flux, markedly reduced ROS levels in *Bmi1*$^{-/-}$ cells. Consistent with the observed rise in Duox1 and Duox2 expression (Fig. 1c), we also observed a small but non-significant decrease in ROS levels after treatment with the NADPH oxidase inhibitor DPI (Fig. 1i). Finally, the use of the mitochondrial redox fluorophore MitoSox Red provided additional support for the hypothesis that the mitochondria are the major source of increased ROS levels observed in *Bmi1*$^{-/-}$ cells (Fig. 1j and Supplementary Fig. 3).

Antioxidants can rescue *Bmi1*$^{-/-}$ mice

We next sought to assess whether the increase in ROS levels might contribute to the various *in vivo* phenotypic defects observed in *Bmi1*$^{-/-}$ mice. We randomized four-week-old *Bmi1*$^{-/-}$ mice to treatment with and without the antioxidant scavenger *N*-acetylcysteine (NAC). After one week of treatment, ROS levels in *Bmi1*$^{-/-}$ thymocytes had been reduced to near wild-type levels (Fig. 2a). Coincident with this reduction, we noted a marked increase in the overall size of the thymus in the antioxidant-treated *Bmi1*$^{-/-}$ mice (Fig. 2b). This

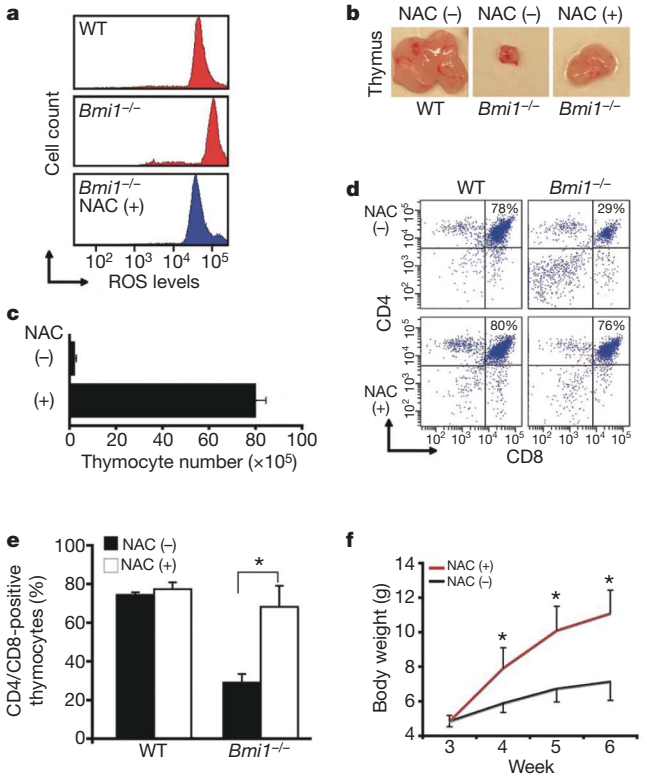


Figure 2 | Antioxidant treatment rescues *Bmi1*$^{-/-}$ thymocytes. **a**, ROS in thymocytes isolated from four-week-old wild-type (WT) mice or *Bmi1*$^{-/-}$ mice randomized to NAC treatment for one week before collection. **b**, Overall thymus size in 4-week-old mice either treated for one week with NAC (+) of left untreated (−). **c**, Total number of thymocytes recovered in 4-week-old *Bmi1*$^{-/-}$ mice either treated for one week with NAC or left untreated (mean and s.d.; $n = 4$ animals per group). **d**, Representative assessment of thymocyte maturation ($CD4^+CD8^+$ cells) in 4-week-old WT or *Bmi1*$^{-/-}$ mice that were treated with NAC for one week or left untreated before collection. **e**, Quantitative assessment of thymocyte maturation in 4-week-old mice randomized for one week of antioxidant treatment (mean and s.d.; $n = 4$ animals per group). **f**, Weight of *Bmi1*$^{-/-}$ mice with or without antioxidant treatment beginning after weaning (mean and s.d.; $n = 5$–7 animals per group). $^*P < 0.05$.

increase in thymus size was also accompanied by a significant increase in the overall number of total thymocytes in the antioxidant-treated *Bmi1*$^{-/-}$ animals, although *Bmi1*$^{-/-}$ mice treated with NAC still had reduced number of thymocytes when compared to wild-type animals (Fig. 2c). Previous studies have also demonstrated that *Bmi1*$^{-/-}$ mice also have defects in thymocyte maturation[3,4]. As noted in Fig. 2d, approximately 80% of cells isolated from wild-type mice were CD4$^+$CD8$^+$ double-positive thymocytes. A one-week administration of NAC to wild-type mice did not appreciably alter this distribution. In contrast, the frequency of CD4$^+$CD8$^+$ double-positive thymocytes observed in untreated *Bmi1*$^{-/-}$ mice was markedly reduced, yet normalized after NAC administration (Fig. 2d, e). Finally, NAC administration had effects beyond altering thymocyte function, because this intervention also appeared to partially correct the failure-to-thrive phenotype seen in *Bmi1*$^{-/-}$ animals (Fig. 2f).

Quantitative real-time polymerase chain reaction (rtPCR) analysis demonstrated that the improvement seen in *Bmi1*$^{-/-}$ mice after antioxidant treatment was not accompanied by any apparent alteration in *Ink4a/Arf* transcription (Fig. 3a). Thus, the Bmi1-mediated rise in ROS does not seem to be required for the subsequent induction of Ink4a/Arf. There is evidence that, in certain situations, a rise in p16^{Ink4a} expression can be associated with, and appears to be required for, a subsequent increase in oxidant levels[15]. Thus, it seemed possible that NAC was rescuing *Bmi1*$^{-/-}$ mice by scavenging oxidants that were produced as a result of p16^{Ink4a} expression. To test this possibility formally, we analysed the levels of ROS in thymocytes obtained from *Bmi1*$^{-/-}$ mice or from mice doubly deficient in both *Bmi1* and *p16*^{Ink4a} (Fig. 3b and Supplementary Fig. 4). The induction of p16^{Ink4a} does not seem to be required for the rise in ROS levels observed in *Bmi1*$^{-/-}$ cells.

Bmi1 regulates the DDR pathway

To understand which other pathways might be activated by the observed increase in ROS levels, we took advantage of previous observations demonstrating that oxidative stress can trigger activation of the DDR pathway[16]. As noted in Fig. 3c, *Bmi1*$^{-/-}$ thymocytes had increased levels of 8-oxoguanine, a stable marker of oxidatively damaged DNA. In addition, consistent with ongoing DNA damage, we observed increased nuclear foci of 53BP1 (also known as Trp53bp1), another known hallmark of DNA damage and DDR activation[17]. In *Bmi1*$^{-/-}$ mice treated with NAC for one week before collection, the number of 53BP1 nuclear foci was markedly reduced (Fig. 3d). Other DDR components including Chk2 were also activated in *Bmi1*$^{-/-}$ cells, and again this activation was reduced by antioxidant treatment (Fig. 3e). These results indicate that the sustained levels of ROS seen in *Bmi1*$^{-/-}$ mice are sufficient to damage DNA directly and to engage the DDR pathway. To underscore this point further, we observed that treatment of thymocytes with exogenous hydrogen peroxide could activate the DDR pathway and that cells obtained from *Chk2*$^{-/-}$ mice were largely protected from ROS-induced cell death (Supplementary Fig. 5). These results indicated that interruption of the DDR pathway by deletion of Chk2 may provide some benefit to the oxidatively stressed *Bmi1*$^{-/-}$ cells and tissues. To begin to address this hypothesis formally, we assessed the survival in culture of wild-type, *Chk2*$^{-/-}$, *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ thymocytes. Whereas *Bmi1*$^{-/-}$ thymocytes rapidly lost viability in culture, the survival of thymocytes lacking both Bmi1 and Chk2 was similar to that in wild-type cells (Fig. 3f). A similar rescue of *Bmi1*$^{-/-}$ thymocytes could also be obtained by supplementing the media with NAC (Supplementary Fig. 6).

Chk2 deletion rescues *Bmi1*$^{-/-}$ mice

To understand further how the activation of the DDR pathway might contribute to the overall *Bmi1*$^{-/-}$ phenotype, we asked whether mice lacking both Bmi1 and Chk2 demonstrated any phenotypic improvement compared to *Bmi1*$^{-/-}$ mice. Previous results have demonstrated that, whereas Chk2-deficient cells have an increase in radioresistance, *Chk2*$^{-/-}$ mice do not exhibit a high rate of spontaneous tumour formation and appear, for the most part, to be phenotypically identical to wild-type mice[18,19]. We first examined *in vivo* thymocyte maturation in either *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ animals. As previously noted with antioxidant treatment, the overall size of the thymus, as well as the spleen, was significantly larger in the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice than in *Bmi1*$^{-/-}$-only mice (Supplementary Fig. 7). Similarly, the previously observed reduction in the number of CD4$^+$CD8$^+$ double-positive thymocytes

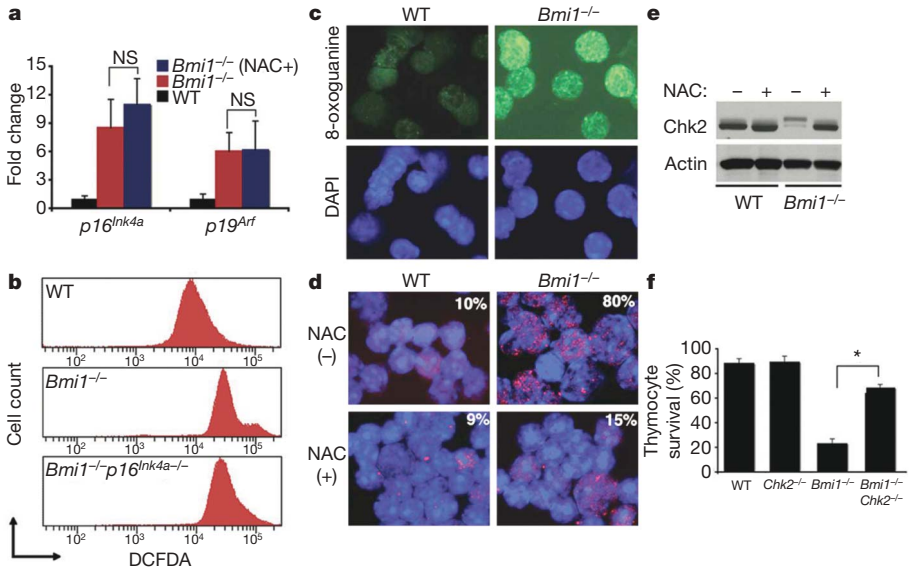


Figure 3 | Activation of the DDR pathway in *Bmi1*$^{-/-}$ thymocytes occurs through a redox-sensitive pathway. **a**, Quantitative rtPCR analysis of *Ink4a/Arf* expression in thymocytes obtained from mice randomized to antioxidant therapy (mean and s.d.; $n = 3$ animals per group). NS, not significant. **b**, ROS levels in thymocytes obtained from WT, *Bmi1*$^{-/-}$ or combined *Bmi1*/*p16*^{Ink4a}-deleted mice. **c**, Levels of the oxidatively modified nucleotide 8-oxoguanine in isolated thymocytes. DAPI (4,6-diamidino-2-phenylindole) staining (blue) was used to visualize nuclei. **d**, 53BP1 nuclear foci thymocytes obtained from mice treated for one week before collection with the antioxidant NAC or left untreated. Overall percentage of thymocytes demonstrating activation of the DDR in each condition is shown. **e**, Western blot analysis for Chk2 activation in thymocytes obtained from WT or *Bmi1*$^{-/-}$ mice. **f**, Assessment of primary thymocyte survival in culture 24 h after isolation.

in *Bmi1*$^{-/-}$ mice was significantly less marked in mice lacking both *Bmi1* and *Chk2* (Fig. 4a). We also noted significant alterations in the overall architecture of the thymus. In contrast to wild-type mice, *Bmi1*$^{-/-}$ mice lacked a distinct separation between thymic cortex and medulla (Fig. 4b). A similar change in overall thymus architecture has been noted in other animal models with impaired thymocyte maturation[20,21]. In contrast, a near normal compartmentalization of thymic cortex and medulla was observed in the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice (Fig. 4b). Given that *Chk2* deletion seems to protect thymocytes from ROS-mediated cell death (Fig. 3f and Supplementary Fig. 5), we postulated that the absence of Chk2 could be rescuing thymocyte number and maturation by inhibiting apoptosis. Consistent with this notion, levels of apoptosis were increased in the cortex of *Bmi1*$^{-/-}$ compared to wild-type animals, whereas this level of overall apoptosis was reduced in animals deficient in both *Bmi1* and *Chk2* (Fig. 4c). As previously observed with antioxidant treatment, the improvement observed in the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice did not seem to be the result of alterations in the induction of the *Ink4a/Arf* locus (Fig. 4d and Supplementary Fig. 8). Similarly, deletion of *Chk2* did not directly alter the level of ROS induced by the absence of Bmi1 expression (Supplementary Fig. 9). Interestingly, the combination of NAC treatment and *Chk2* deletion was slightly more effective at rescuing *Bmi1*$^{-/-}$ thymocyte number and maturation than *Chk2* deletion alone, indicating that some of the effects of ROS may be independent of the DDR response (Supplementary Fig. 10).

We next assessed whether the absence of Chk2 could also improve the number or functional capacity of *Bmi1*$^{-/-}$ haematopoietic stem or progenitor cells. To begin to address this issue, we measured the number of LSK cells in either *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice. This cell population is generally believed to contain abundant progenitor cells, as well as rarer, long-term repopulating cells. As noted in Fig. 4e, the number of LSK cells in mice lacking both *Bmi1* and *Chk2* was significantly higher than that seen in *Bmi1*$^{-/-}$ animals. We next sought to assess the functional capacity of the progenitor cells isolated from *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice. Bone marrow cells from *Bmi1*$^{-/-}$ mice were markedly impaired in their *in vitro* colony-forming ability whereas the colony-forming capacity of *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ bone marrow cells was significantly improved (Supplementary Fig. 11). In a further *in vivo* test of progenitor function, we measured spleen colony-forming units (CFU-S) that formed after injection of total bone marrow (1×10^5 cells) into irradiated hosts. Again, using this *in vivo* test of progenitor cell function, we noted a marked impairment of *Bmi1*$^{-/-}$ bone marrow that was significantly less pronounced in the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice (Fig. 4f).

Although these results indicate that *Chk2* deletion results in improvement in the number and function of *Bmi1*$^{-/-}$ progenitor cells, they do not address whether there is also a concomitant improvement in long-term repopulating ability. To test this, we performed competitive repopulation studies using wild-type, *Bmi1*$^{-/-}$, *Bmi1*$^{+/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ bone marrow. As previously reported[1,8], consistent with a defect in stem cell maintenance and self-renewal, we observed that *Bmi1*$^{-/-}$ bone marrow was unable to contribute in such long-term repopulation assays (Supplementary Fig. 12). We also observed essentially no contribution to long-term repopulation from the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ bone marrow (Fig. 4g and Supplementary Fig. 12). We therefore conclude that, whereas *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice have an increased number of LSK cells and an apparent improvement in progenitor cell function, deletion of *Chk2* does not rescue the self-renewal defect seen in *Bmi1*$^{-/-}$ mice.

The improvement in *Bmi1*$^{-/-}$ mice that was observed by concomitant deletion of *Chk2* extended beyond thymocytes and haematopoietic progenitors. Indeed, the overall appearance of these doubly

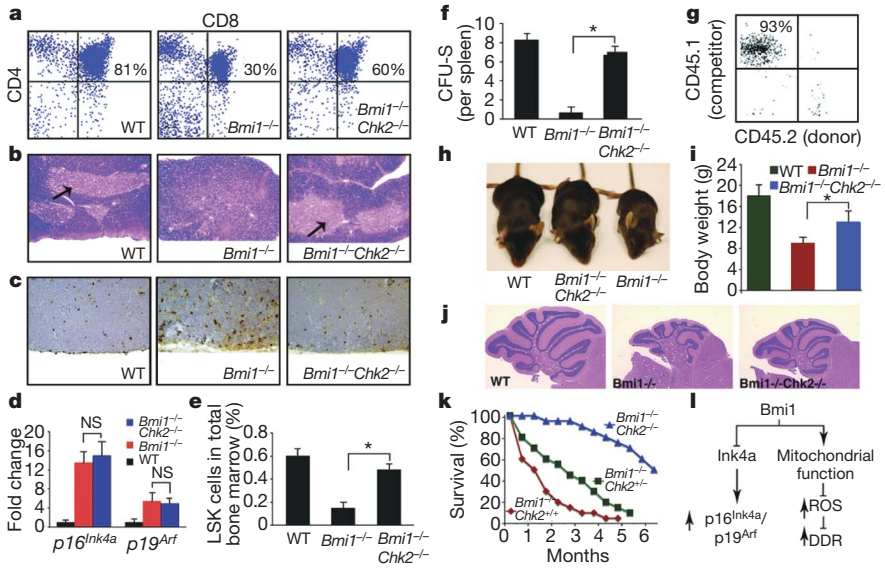


Figure 4 | Inhibition of the DDR pathway by Chk2 deletion rescues multiple defects in *Bmi1*$^{-/-}$ mice. **a**, *Chk2* deletion restores *in vivo* thymocyte maturation of *Bmi1*$^{-/-}$ mice. WT, wild type. **b**, Architecture of the thymus in WT, *Bmi1*$^{-/-}$ and combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice. The arrows point to a normal medulla region. **c**, Representative TUNEL staining in the cortex region of the thymus. **d**, Quantitative rtPCR analysis of *Ink4a/Arf* induction in *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ thymocytes (mean and s.d.; $n = 3$ animals per group). **e**, Frequency of LSK cells in total bone marrow (*$P < 0.01$; mean and s.d., $n = 6$ animals per group). **f**, *In vivo* CFU-S in lethally irradiated mice infused with equal numbers of bone marrow cells obtained from WT, *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice (*$P < 0.01$; mean and s.d., $n = 5$ animals per group). **g**, Peripheral blood composition four months after competitive repopulation in which equal amounts of WT competitor bone marrow cells (CD45.1) and combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ donor bone marrow (CD45.2) were transplanted into an irradiated host (CD45.2). We observed little to no contribution of the donor-derived *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ cells after such competitive repopulation experiments. **h**, Representative appearance of WT, *Bmi1*$^{-/-}$ or combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice. **i**, Analysis of body weight at 6 weeks of age (*$P < 0.01$; mean and s.d.; $n = 6$ per group). **j**, Defects in cerebellar architecture observed in *Bmi1*$^{-/-}$ mice are rescued by Chk2 deletion. **k**, Lifespan of *Bmi1*$^{-/-}$ mice with varying Chk2 status. $P < 0.05$ *Chk2*$^{+/-}$ versus *Chk2*$^{+/+}$ and $P < 0.001$ for *Chk2*$^{-/-}$ versus *Chk2*$^{+/+}$; $n \geq 15$ animals per genotype. **l**, The Polycomb protein Bmi1 normally simultaneously represses the *Ink4a/Arf* locus leading to reduced p16^{Ink4a} and p19^{Arf} expression as well as modulating mitochondrial function to lower ROS levels and suppress activation of the DDR pathway. Activation of both Ink4a/Arf and the DDR have been separately linked to tumour suppression and stem cell ageing.

deficient mice was improved compared to the *Bmi1*$^{-/-}$ animals (Fig. 4h). This was also accompanied by a significant increase in overall body weight (Fig. 4i). In addition, whereas the *Bmi1*$^{-/-}$ mice had a severe ataxia, this phenotype also appeared to be mitigated in the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ animals. Consistent with such improvement, the cerebellum of the double-deficient mice showed a marked increase in cellularity and was more developed than what was observed in *Bmi1*$^{-/-}$ mice (Fig. 4j). Similarly, whereas *Bmi1*$^{-/-}$ mice were infertile, both male and female combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice were capable of generating viable offspring. In contrast, the deletion of *Chk2* did not rescue previously described changes in the axial skeleton, because we observed that approximately 80% of the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice had 6 as opposed to the normal 7 vertebrosternal ribs[3].

Finally, we noted a number of progeriod features in the *Bmi1*$^{-/-}$ mice that were largely absent in age-matched *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice (Supplementary Fig. 13). On the basis of these observations, we asked whether the absence of Chk2 could extend the median and maximum lifespan of animals lacking Bmi1. In our colony, the median survival of *Bmi1*$^{-/-}$ mice was roughly 1 month and essentially all mice succumbed by 3–4 months (Fig. 4k). Deletion of one copy of *Chk2* provided a survival advantage, whereas deletion of both copies of *Chk2* was even more effective and extended the median survival of the *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ mice to approximately 6 months.

Conclusions

We demonstrate that the absence of Bmi1 leads to increased expression of a collection of gene products involved in mitochondrial function and ROS homeostasis. Cells lacking Bmi1 have significant mitochondrial dysfunction accompanied by a sustained increase in ROS that is sufficient to engage the DDR pathway. Treatment with an antioxidant or interruption of the DDR by *Chk2* deletion substantially improves some, but not all, aspects of the *Bmi1*$^{-/-}$ mice. Our observations further indicate that, in *Bmi1*$^{-/-}$ mice, ROS act independently of the Ink4a/Arf pathway. In addition, although in the context of the *Bmi1*$^{-/-}$ mouse deletion of the *Ink4a/Arf* locus results in substantial rescue of the LT-HSC defect, it is interesting to note that the beneficial effects of *Chk2* deletion are exclusively confined to very early haematopoietic progenitor cells.

Activation of the DDR response has been invoked as a natural consequence of age-associated damage. For instance, when compared to younger HSCs, purified HSCs from old mice demonstrate increased levels of γ-H2AX nuclear foci—a marker of DNA damage[22]. Although such old HSCs do not decline in absolute number, the activation of the DDR does correlate with impairment of stem cell function. Similarly, in a number of mouse genetic models in which there are inherited defects in various components of the DNA repair apparatus, there is also evidence for a reduction in functional capacity of transplanted HSCs[23,24]. In addition, recent evidence indicates that the self-renewal of cancer stem cells can also be limited by oncogene-induced DNA damage[25]. In other animal models, the age-dependent increase in p16^{Ink4a} expression has been linked to age-dependent decline in stem cell function[26]. Furthermore, in several independent models, deletion of *p16*^{Ink4a} leads to improved functional capacity of neural, pancreatic and haematopoietic stem cells[27–29]. Thus, both the DDR pathway and a rise in p16^{Ink4a} have been linked to stem and progenitor cell ageing and dysfunction. Outside of the stem and progenitor compartment, both activation of the Ink4a/Arf locus and engagement of the DDR pathway have also been separately implicated in mediating cellular senescence, growth arrest and/or cell death induced by various stresses[30,31]. Activation of these two pathways has been invoked as an important tumour suppressor barrier, because both pathways act as potent inhibitors of the proliferation or propagation of damaged cells. Interestingly, our data indicate that the absence of Bmi1 results in the simultaneous engagement of both of these stress-activated and tumour-suppressive pathways and that both of these pathways appear to separately contribute to the overall phenotype of the *Bmi1*$^{-/-}$ mice (Fig. 4l). Furthermore, the necessity of stem cells to persist throughout the entire lifespan of the organism indicates that these cells might have particularly stringent requirements to limit ROS production. Our results imply that the Polycomb family of proteins might be uniquely positioned to coordinately regulate mitochondrial function and redox homeostasis with overall stem cell biology.

METHODS SUMMARY

Bmi1$^{+/-}$ and *Chk2*$^{+/-}$ mice have been described previously[3,18]. For the *in vivo* administration of NAC, we randomized animals to normal drinking water, or water containing NAC at 1 mg ml^{-1}, as previously described[10]. For analysis of intracellular ROS, bone marrow cells enriched for LT-HSCs, thymocytes and other cell types were incubated with 5 μM dichlorofluorescein diacetate (DCFDA, Invitrogen) and incubated in a shaker at 37 °C for 30 min, followed immediately by flow cytometry analysis using a LSRII instrument (Becton-Dickinson). Functional mitochondria from mouse tissues were isolated by differential centrifugation while measurement of intact cellular respiration was performed using the Seahorse XF24 analyser[32]. ATP was measured using the ATP-determination kit (Molecular Probes) with 0.2 μg of protein lysate. For competitive repopulation experiments, lethally irradiated C57BJ/6 mice (B6-CD45.2) were competitively reconstituted with 1×10^6 total bone morrow cells from *Bmi1*$^{+/+}$, *Bmi1*$^{+/-}$, *Bmi1*$^{-/-}$ or combined Bmi1/Chk2-deficient mice. Quantitative rtPCR for *Ink4a/Arf* expression or for the expression of other *Bmi1*-regulated genes was performed on an MxP3005P real-time PCR system (Stratagene) using SYBR Green PCR Mastermix (Applied Biosystems) according to the manufacturer's instructions. For assessment of DDR pathway activation, we used freshly isolated thymocytes from 4-week-old *Bmi1*$^{+/+}$ and *Bmi1*$^{-/-}$ mice treated with NAC for one week or left untreated. Cells were analysed for nuclear 53BP1 (Novus) foci or activation of Chk2 (BD Transduction Laboratories). The oxidatively modified base 8-oxoguanine was detected by avidin–FITC (fluorescein isothiocyanate) staining of fixed cells as described previously[33,34].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 March; accepted 8 April 2009.
Published online 29 April 2009.

1. Park, I. K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423,** 302–305 (2003).
2. Molofsky, A. V. *et al.* Bmi-1 dependence distinguishes neural stem cell self-renewal from progenitor proliferation. *Nature* **425,** 962–967 (2003).
3. van der Lugt, N. M. *et al.* Posterior transformation, neurological abnormalities, and severe hematopoietic defects in mice with a targeted deletion of the *bmi-1* proto-oncogene. *Genes Dev.* **8,** 757–769 (1994).
4. Jacobs, J. J., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene *bmi-1* regulates cell proliferation and senescence through the *ink4a* locus. *Nature* **397,** 164–168 (1999).
5. Bruggeman, S. W. *et al.* Ink4a and Arf differentially affect cell proliferation and neural stem cell self-renewal in Bmi1-deficient mice. *Genes Dev.* **19,** 1438–1443 (2005).
6. Molofsky, A. V., He, S., Bydon, M., Morrison, S. J. & Pardal, R. Bmi-1 promotes neural stem cell self-renewal and neural development but not mouse growth and survival by repressing the p16Ink4a and p19Arf senescence pathways. *Genes Dev.* **19,** 1432–1437 (2005).
7. Bracken, A. P. *et al.* The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. *Genes Dev.* **21,** 525–530 (2007).
8. Oguro, H. *et al.* Differential impact of Ink4a and Arf on hematopoietic stem cells and their bone marrow microenvironment in Bmi1-deficient mice. *J. Exp. Med.* **203,** 2247–2253 (2006).
9. Tothova, Z. *et al.* FoxOs are critical mediators of hematopoietic stem cell resistance to physiologic oxidative stress. *Cell* **128,** 325–339 (2007).
10. Ito, K. *et al.* Regulation of oxidative stress by ATM is required for self-renewal of haematopoietic stem cells. *Nature* **431,** 997–1002 (2004).
11. Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121,** 1109–1121 (2005).
12. Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H. & Helin, K. Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.* **20,** 1123–1136 (2006).
13. Fasano, C. A. *et al.* shRNA knockdown of Bmi-1 reveals a critical role for p21-Rb pathway in NSC self-renewal during development. *Cell Stem Cell* **1,** 87–99 (2007).
14. Balaban, R. S., Nemoto, S. & Finkel, T. Mitochondria, oxidants, and aging. *Cell* **120,** 483–495 (2005).
15. Takahashi, A. *et al.* Mitogenic signalling and the p16INK4a-Rb pathway cooperate to enforce irreversible cellular senescence. *Nature Cell Biol.* **8,** 1291–1297 (2006).

16. Lombard, D. B. *et al.* DNA repair, genome stability, and aging. *Cell* **120,** 497–512 (2005).
17. Adams, M. M. & Carpenter, P. B. Tying the loose ends together in DNA double strand break repair with 53BP1. *Cell Div* **1,** 19 (2006).
18. Takai, H. *et al.* Chk2-deficient mice exhibit radioresistance and defective p53-mediated transcription. *EMBO J.* **21,** 5195–5205 (2002).
19. Hirao, A. *et al.* Chk2 is a tumor suppressor that regulates apoptosis in both an ataxia telangiectasia mutated (ATM)-dependent and an ATM-independent manner. *Mol. Cell. Biol.* **22,** 6521–6532 (2002).
20. Okada, H. *et al.* Survivin loss in thymocytes triggers p53-mediated growth arrest and p53-independent cell death. *J. Exp. Med.* **199,** 399–410 (2004).
21. Zaugg, K. *et al.* Cross-talk between Chk1 and Chk2 in double-mutant thymocytes. *Proc. Natl Acad. Sci. USA* **104,** 3805–3810 (2007).
22. Rossi, D. J. *et al.* Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature* **447,** 725–729 (2007).
23. Reese, J. S., Liu, L. & Gerson, S. L. Repopulating defect of mismatch repair-deficient hematopoietic stem cells. *Blood* **102,** 1626–1633 (2003).
24. Nijnik, A. *et al.* DNA repair is limiting for haematopoietic stem cells during ageing. *Nature* **447,** 686–690 (2007).
25. Viale, A. *et al.* Cell-cycle restriction limits DNA damage and maintains self-renewal of leukaemia stem cells. *Nature* **457,** 51–56 (2009).
26. Rossi, D. J., Jamieson, C. H. & Weissman, I. L. Stems cells and the pathways to aging and cancer. *Cell* **132,** 681–696 (2008).
27. Janzen, V. *et al.* Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16INK4a. *Nature* **443,** 421–426 (2006).
28. Molofsky, A. V. *et al.* Increasing p16INK4a expression decreases forebrain progenitors and neurogenesis during ageing. *Nature* **443,** 448–452 (2006).
29. Krishnamurthy, J. *et al.* p16INK4a induces an age-dependent decline in islet regenerative potential. *Nature* **443,** 453–457 (2006).
30. Collado, M., Blasco, M. A. & Serrano, M. Cellular senescence in cancer and aging. *Cell* **130,** 223–233 (2007).
31. Finkel, T., Serrano, M. & Blasco, M. A. The common biology of cancer and ageing. *Nature* **448,** 767–774 (2007).
32. Ferrick, D. A., Neilson, A. & Beeson, C. Advances in measuring cellular bioenergetics using extracellular flux. *Drug Discov. Today* **13,** 268–274 (2008).
33. Struthers, L., Patel, R., Clark, J. & Thomas, S. Direct detection of 8-oxodeoxyguanosine and 8-oxoguanine by avidin and its analogues. *Anal. Biochem.* **255,** 20–31 (1998).
34. Neumann, C. A. *et al.* Essential role for the peroxiredoxin Prdx1 in erythrocyte antioxidant defence and tumour suppression. *Nature* **424,** 561–565 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to M. Clarke for providing the initial supply of *Bmi1*$^{-/-}$ mice, to J. Moss for the gift of anti-PARP antibodies and to M. Daniels and the NHLBI electron microscope core for their assistance. This work was supported by funding from the NIH intramural program and the Ellison Medical Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.F. (finkelt@nih.gov) or L.C. (Liu.Cao@nih.gov).

METHODS

Cells and mice. *Bmi1*$^{+/-}$ and *Chk2*$^{+/-}$ mice have been described previously[3,18]. *Bmi1*$^{+/-}$ C57BL/6 mice and *Chk2*$^{+/-}$ mice on a mixed 129/Black Swiss genetic background were crossed to generate *Bmi1*$^{+/+}$ and *Bmi1*$^{-/-}$ mice with various Chk2 status. Littermate controls were used for analysis and all mice were genotyped routinely by PCR with mouse tail DNA. Except where indicated, histological and biological analysis was performed for both male and female mice at 1–3 months of age. Mice deleted specifically for *p16*^{Ink4a} have been described[35] and were obtained from the National Cancer Institute (MMHCC repository, Fredrick MD). For the *in vivo* administration of NAC, we randomized animals to normal drinking water, or water containing NAC at 1 mg ml^{-1} as previously described[10]. In addition, owing to the overall frailty of *Bmi1*$^{-/-}$ mice, the food of animals treated or untreated with antioxidant was soaked in NAC-treated or -untreated drinking water and placed inside the cage.

Intracellular ROS analysis. Total bone marrow cells, LSK cells and cells stained for SLAM family members (CD150 and CD48; BioLegend) as well as CD41 (BioLegend) were isolated as previously described[11,36]. In general, thymocytes and splenocytes were isolated from 6–8-week-old mice and were plated at an initial density of 1×10^6 cells per ml in Dulbecco Modified Essential Medium (DMEM) supplemented with 10% fetal bovine serum. Thymocyte populations were analysed where indicated by CD4 and CD8 staining (BD Biosciences). For analysis of intracellular ROS, bone marrow cells enriched for LT-HSCs, thymocytes and other cell types were incubated with 5 μM DCFDA (Invitrogen) and placed in a shaker at 37 °C for 30 min, followed immediately by flow cytometry analysis using an LSRII instrument (Becton-Dickinson). Viable cells (1×10^6) as determined by 7-AAD staining were used for the analysis of ROS levels. Where indicated, cells were treated with either rotenone, carbonyl cyanide p-trifluormethoxyphenylhydrazone (FCCP) or diphenyleneiodonium chloride (DPI) for 10 min before assessment of ROS levels.

Mitochondrial isolation, functional studies and ATP determination. For isolation of functional mitochondria from mouse tissues, we rapidly collected, washed and minced mouse hearts in ice-cold Isolation Buffer I containing 210 mM mannitol, 70 mM sucrose, 5 mM HEPES (pH 7.4), 1 mM EGTA and 0.5 mg ml^{-1} BSA. The hearts were then homogenized in this buffer with a glass-Teflon motorized homogenizer. The mitochondrial fraction was isolated by differential centrifugation and subsequently washed twice before resuspension in Isolation Buffer I at a concentration of 0.5 mg ml^{-1} before functional assessment.

Mitochondrial respiration from intact mitochondria was measured by standard protocols using a respiration buffer containing 225 mM mannitol, 75 mM sucrose, 10 mM KCl, 10 mM Tris-HCl and 5 mM KH_2PO_4 at pH 7.2. Glutamate (5 mM), malate (5 mM) and ADP (1 mM) were used to assay respiration through complex I. Succinate (5 mM), rotenone (1 μM) and ADP (1 mM) were used to assay complex II respiration. Respiratory control ratios (RCRs; state 3/state 4) were determined after inhibition of mitochondrial ATPase with oligomycin. Measurement of intact cellular respiration was performed using the Seahorse XF24 analyser[32]. Respiration was measured under basal condition, in the presence of the mitochondrial inhibitor oligomycin (0.5 μM) and in the presence of the mitochondrial uncoupler FCCP (1 μM) to assess maximal oxidative capacity as we have described previously[37].

ATP was measured using the ATP-determination kit (Molecular Probes). For *in vivo* measurements, mice tissues were rapidly collected and immediately placed in ice-cold ATP buffer (20 mM Tris, pH 7.5, 0.5% Nonidet P-40, 25 mM NaCl, 2.5 mM EDTA) for 5 min. Tissue samples then underwent one five-second round of sonication. Lysates were than centrifuged at 13,000*g* for 30 min and the supernatant measured for protein concentration. ATP concentration was measured from 0.2 μg of this protein lysate. We always measured ATP from freshly isolated (not frozen) tissues.

HSC cultures and colony-forming assays. CFU-GM *in vitro* assays were performed according to the manufacturer's instructions (StemCell Technologies). In brief, freshly isolated total bone marrow cells were cultured in semisolid MethoCult media supplemented with 10 ng ml^{-1} GM-CSF. Cells were plated in triplicate at an initial density of 2×10^4 cells per well of a 12-well plate. After 5–7 days of culture, CFU-GM colonies were assessed. The CFU-S assay was performed as previously described[38]. In brief, 1×10^5 bone marrow cells from five donors of each group (*Bmi1*$^{-/-}$, combined *Bmi1*$^{-/-}$*Chk2*$^{-/-}$ and wild-type controls) were transplanted into each of three 12-week-old normal B6 recipients pre-irradiated at 950 rad. Six B6 mice that received the same dose of irradiation and no subsequent bone marrow transplantation were used as negative controls. The spleens of all animals were fixed on day 12 after transplantation, and the total numbers of colonies per spleen were assessed.

For competitive repopulation experiments, lethally irradiated C57BJ/6 mice (B6-CD45.2) were competitively reconstituted with 1×10^6 total bone morrow cells from *Bmi1*$^{+/+}$, *Bmi1*$^{+/-}$, *Bmi1*$^{-/-}$ or combined Bmi1/Chk2-deficient mice. We used five recipients in each group and each recipient also received competitive repopulation with 1×10^6 bone marrow cells obtained from C57BJ/6 mice congenic for the CD45 locus (B6-CD45.1). Analysis was performed on a monthly basis after transplantation up to month four. Peripheral blood cells of the recipients were stained with antibodies against CD45.1, CD45.2, B220, CD3 and Mac-1 to monitor the reconstitution of both myeloid and lymphoid lineages.

Gene expression. Total RNA was isolated by using Trizol reagent and subsequently treated with DNase I (Invitrogen). Complementary DNA was prepared using Taqman reverse transcription reagents and oligo-dT primers. Quantitative rtPCR for *Ink4a/Arf* expression or for the expression of other Bmi1-regulated genes was performed with 50 ng cDNA on an MxP3005P real-time PCR system (Stratagene) using SYBR Green PCR Mastermix (Applied Biosystems) according to the manufacturer's instructions. For primer sequences, see Supplementary Information and Supplementary Methods.

35. Sharpless, N. E. *et al.* Loss of p16Ink4a with retention of p19Arf predisposes mice to tumorigenesis. *Nature* **413,** 86–91 (2001).
36. Smith, A. L., Ellison, F. M., McCoy, J. P. Jr & Chen, J. c-Kit expression and stem cell factor-induced hematopoietic cell proliferation are up-regulated in aged B6D2F1 mice. *J. Gerontol. A Biol. Sci. Med. Sci.* **60,** 448–456 (2005).
37. Schieke, S. M. *et al.* The mammalian target of rapamycin (mTOR) pathway regulates mitochondrial oxygen consumption and oxidative capacity. *J. Biol. Chem.* **281,** 27643–27652 (2006).
38. TeKippe, M., Harrison, D. E. & Chen, J. Expansion of hematopoietic stem cell phenotype and activity in Trp53-null mice. *Exp. Hematol.* **31,** 521–527 (2003).

ARTICLES

The origin of the electrostatic perturbation in acetoacetate decarboxylase

Meng-Chiao Ho[1], Jean-François Ménétret[1], Hiro Tsuruta[2] & Karen N. Allen[1]†

Acetoacetate decarboxylase (AADase) has long been cited as the prototypical example of the marked shifts in the pK_a values of ionizable groups that can occur in an enzyme active site. In 1966, it was hypothesized that in AADase the origin of the large pK_a perturbation (−4.5 log units) observed in the nucleophilic Lys 115 results from the proximity of Lys 116, marking the first proposal of microenvironment effects in enzymology. The electrostatic perturbation hypothesis has been demonstrated in a number of enzymes, but never for the enzyme that inspired its conception, owing to the lack of a three-dimensional structure. Here we present the X-ray crystal structures of AADase and of the enamine adduct with the substrate analogue 2,4-pentanedione. Surprisingly, the shift of the pK_a of Lys 115 is not due to the proximity of Lys 116, the side chain of which is oriented away from the active site. Instead, Lys 116 participates in the structural anchoring of Lys 115 in a long, hydrophobic funnel provided by the novel fold of the enzyme. Thus, AADase perturbs the pK_a of the nucleophile by means of a desolvation effect by placement of the side chain into the protein core while enforcing the proximity of polar residues, which facilitate decarboxylation through electrostatic and steric effects.

AADase, a 365 kDa homododecameric enzyme that catalyses the conversion of acetoacetate to acetone, is a key component in the anaerobic metabolism of carbohydrate in solventogenic bacteria. In the early 1960s, Westheimer used AADase to pioneer the application of methods in physical organic chemistry to the study of the chemical and catalytic mechanism of enzymes[1]. These studies revealed that the mechanism of AADase proceeds through a Schiff-base intermediate formed by reaction of Lys 115 with substrate[2–5]. A reporter group used to measure directly the pK_a of Lys 115 revealed it to be 5.96, a value 4.5 orders of magnitude below that expected[6]. Westheimer hypothesized that the pK_a of Lys 115 was electrostatically perturbed by charge–charge repulsion due to the proximity of the protonated ε-amino group of an adjacent Lys 116 (ref. 4). This marked the first appearance of the proposal of microenvironment effects in enzymology, a hypothesis that has since been demonstrated in a number of enzymes[7] but never for the enzyme that inspired its conception, AADase.

Here we present the X-ray crystal structures of AADase from *Clostridium acetobutylicum* (CaAAD) and *Chromobacterium violaceum* (CvAAD, 50% sequence identity) at 2.4 Å and 2.1 Å resolution, respectively. CaAAD ($k_{cat} = 165\,s^{-1}$, $K_m = 4.1$ mM) and CvAAD ($k_{cat} = 349\,s^{-1}$, $K_m = 5.7$ mM) have the same catalytic efficiency, subunit structures (1.2 Å root mean squared deviation, Supplementary Fig. 1) and overall oligomerization properties. The AADase structure exhibits a previously unknown fold (Fig. 1a and Supplementary Fig. 2) consisting predominately of β-strands (15 β-strands and 5 short α-helices with a β-strand and α-helical content of 44.3% and 15.6%, respectively). The tertiary structure is formed by three antiparallel β-sheets, dominated by a central seven-stranded cone-shaped β-barrel (β-cone) and flanked by four-stranded and three-stranded β-sheet structures (Fig. 1b). The strands of the β-cone are twisted such that the first and last strands are perpendicular to one another, similar to a single blade of a β-propeller. The order of the three sheets is discontinuous, with the carboxy-terminal sequence of the protein forming the final strand of all three sheets. Each protomer includes a complete active site, wherein the β-cone encompasses the active site in its hollow core, which extends to a depth of 26.8 Å, with the catalytic Lys 115 positioned at the bottom (CaAAD residue numbering is used throughout). Examination of the solvent-accessible surface shows a single narrow channel (Fig. 1c) leading from bulk solvent at the rim of the cone to the active-site Lys ε-amino moiety. We suggest that the AADase fold be called the Westheimer fold in honour of the many contributions of Frank H. Westheimer to the field of mechanistic enzymology[8].

A hydrophobic active site

The originally proposed pK_a perturbation of Lys 115 by means of Coulombic destabilization by a like-charged residue is not supported by this structure, because the ε-amino groups of Lys 115 and Lys 116 are separated by 14.8 Å (Fig. 2a). Instead, it is the hydrophobic environment of the active site, comprising Phe 26, Leu 71, Tyr 74, Met 96, Leu 98, Tyr 113 and Leu 233 (Fig. 2b), that destabilizes the protonated amine. Furthermore, the side chain of Lys 115 does not form any hydrogen-bonding interactions. Notably, Lys 115 is positioned in parallel with, and in close proximity (~4.7 Å) to, the aromatic ring of Tyr 113, but it is oriented such that there is little potential for stabilization of the lysine ammonium group by cation–π interactions[9]. Indeed, the constrained phi/psi dihedral angles of Pro 114 maintain the relative orientation of Tyr 113 and Lys 115. Sequence alignment demonstrates that this (Y/F)PKK motif and neighbouring hydrophobic residues are conserved in all 25 analogues of AADase (Supplementary Table 2). The structure supports a physical role for Lys 116 in the precise positioning of the nucleophilic Lys 115. This is consistent with previous studies including site-directed mutagenesis and chemical rescue[10] that established the essentiality of the ε-amino group of Lys 116 in catalysis and the maintenance of the depressed pK_a of the nucleophilic Lys. Lys 116 lies at the subunit interface, pinning the β-strand on which it and Lys 115 reside into position by means of

[1]Department of Physiology and Biophysics, Boston University School of Medicine, Boston, Massachusetts 02118-2394, USA. [2]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, MS69, Menlo Park, California 94025-7015, USA. †Present address: Department of Chemistry, Boston University, Boston, Massachusetts 02215-2521, USA.

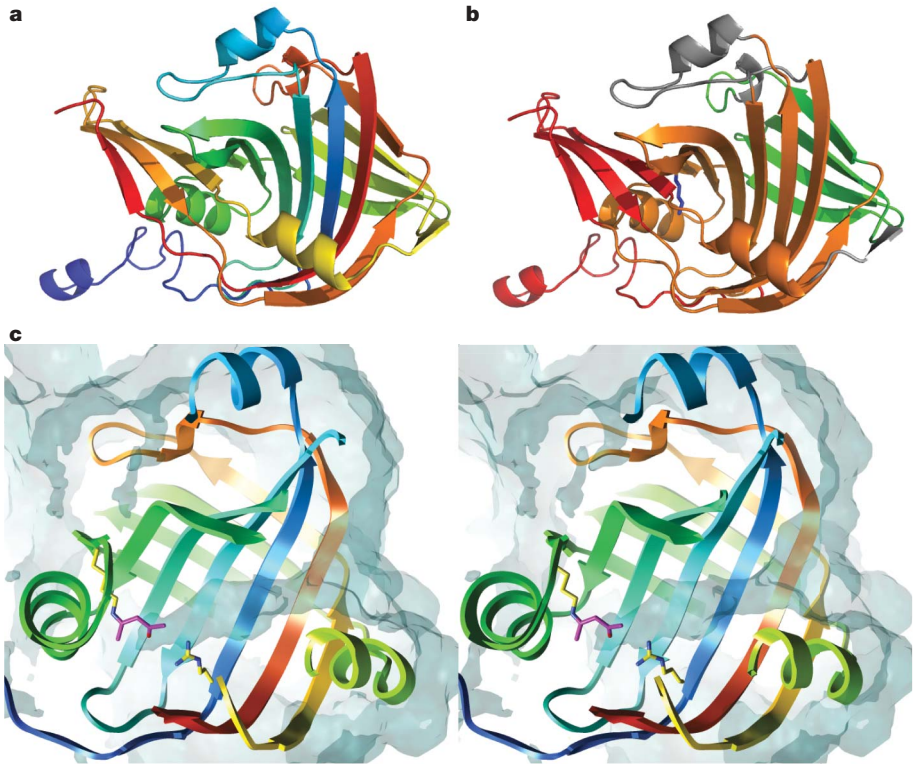


Figure 1 | The AADase tertiary structure depicted as a ribbon diagram. **a**, CaAAD ramped from blue to red (N terminus to C terminus) and, **b**, coloured to differentiate the central cone shaped seven-stranded β-sheet (orange), the four-stranded sheet that comprises the trimer interface (green) and the three-stranded sheet and helix that is a major component of the dimer interface (red). The lysine nucleophile (blue sticks) lies at the bottom of the active site. **c**, Stereo-view of the active site funnel highlighted by means of a Connolly surface (blue transparent surface, showing accessibility to a 1.4 Å radius solvent probe) overlayed on a cutaway of the ribbon diagram coloured as in **a**. For reference, Lys 115 and Arg 29 (cpk coloured ball and stick) from the structure of CvAAD liganded to 2,4-pentanedione (magenta) are shown.

hydrogen-bonding interactions with Ser 16 and the carbonyl oxygen of Met 210 (Fig. 2a).

There are only two charged residues, Arg 29 and Glu 76, near the active-site cavity (Fig. 2b); these are strictly conserved in AADase. Arg 29 and Glu 76 are too distant (6.6 Å and 4.3 Å, respectively) from the Lys 115 ε-amino group to form hydrogen bonds or salt bridges, and thus the local environment is non-polar. Therefore, perturbation of the pK_a of Lys 115 results from the energetically unfavourable process of transferring a charged group from a polar, aqueous solvent (in this case water, with dielectric constant $\varepsilon = 78.54$) to the non-polar interior of a protein (with $\varepsilon = 4$)[11–13]. AADase is not the only example of charge destabilization by means of desolvation in an enzyme; indeed, a similar ΔpK_a of -4.9 was determined for a mutant of staphylococcal nuclease in which a Val buried in the hydrophobic core of the protein was replaced by Lys[14]. Likewise, the reactivity of the Schiff base forming Lys in a catalytic antibody with decarboxylase and aldolase activity at neutral pH was demonstrated to be due to the hydrophobicity of the binding site by means of X-ray crystallography and the analysis of the linear free energy relationship between substrate partitioning into *n*-octanol and k_{cat}/K_m[15]. These cases closely mimic the case of AADase in which the neutral form of the Lys side chain is favoured by placement in a hydrophobic active site. Moreover, binding of a carboxylate group in a site less polar than water has been shown to be an important strategy for catalysis in decarboxylation reactions[16].

The catalytic mechanism

The structure of AADase bound to a ligand was also determined to investigate potential conformational changes upon substrate binding by enzyme and to provide insight into the interactions of the substrate with the active-site residues. The intermediate analogue 2,4-pentanedione acts as a potent inhibitor of AADase ($K_i = 7 \times 10^{-7}$ M for CaAAD)[17] by forming an enamine adduct with the enzyme-based nucleophile, mimicking the substrate Schiff base. The crystal structure of CvAAD complexed with 2,4-pentanedione (2.1 Å resolution) shows no major conformational differences when compared to the unliganded enzyme (0.5 Å root mean squared deviation). The position of the ligand with respect to the solvent-accessible surface (Fig. 1c) shows that the acyl group of the inhibitor sits at the bottom of the active site, with the β-carbonyl group corresponding to the carboxylate of acetoacetate pointing towards the solvent channel. This finding indicates that the funnel shape of the β-cone guides the substrate to the active site. There are only two polar residues in proximity to the bound inhibitor, Arg 29 and Glu 76 (Fig. 2c). The position of Arg 29 in the analogous substrate complex would allow a weak ionic interaction (3.8 Å) with the substrate β-carboxylate, favouring a productive binding mode wherein the carbonyl group of the substrate points towards the catalytic lysine and carboxylate group towards solvent, but without stabilizing the carboxylate ground state. A role in promoting productive substrate binding for Arg 29 is consistent with the results of kinetic analysis of CaAAD bearing an Arg29Gln mutation (Supplementary Table 3). The activity of the mutant was too low to accurately determine the kinetic constants; however, the upper limit for activity shows a diminution of >2,000-fold in maximal rate, which could only be measured at high substrate concentrations (>10 times K_m of wild type). The fact that the catalytic activity of Arg29Gln CaAAD did not increase at pH values above the wild-type optimum for CvAAD of ~5.4 (pH range tested 4.2–6.6, Supplementary Fig. 3) indicates that Arg 29 does not cause the pK_a perturbation of Lys 115.

In the decarboxylation of β-keto acids, the bond undergoing cleavage must be out of the plane of the imine π bond[18]. In the AADase enamine complex, the observed electron density for the 2,4-pentanedione shows that the proximity of Glu 76 forces the group corresponding to the carboxylate group of substrate out of the plane of the enolic oxygen

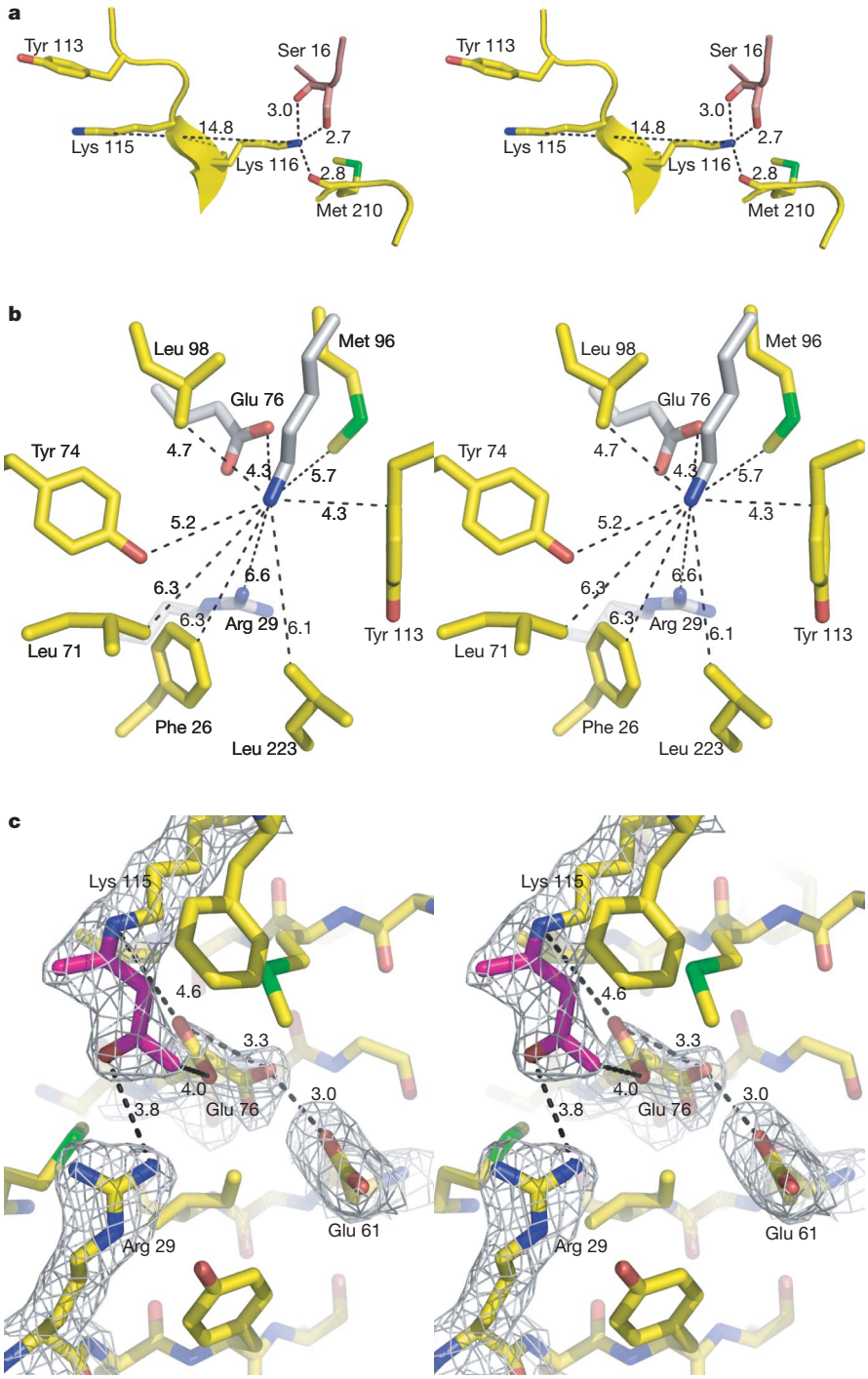


Figure 2 | The positions of Lys 115 and Lys 116 and surrounding environment. **a**, Distances (in Å) between ε-amino groups of Lys 115 and Lys 116 and between Lys 116 and neighbouring residues from the same subunit (yellow) and adjacent subunit (pink) are shown in stereo and labelled. **b**, Stereo view of the active site of AADase with Lys 115 (light blue) and surrounding hydrophobic (yellow) and conserved polar (grey) residues depicted as sticks (distances in Å). **c**, Stereo diagram of the active site of the CvAAD intermediate analogue complex with active site residues (yellow) and 2,4-pentanedione (magenta) shown as sticks overlayed with a $2F_o$–F_c (contoured at 1σ) electron density map (grey cages). Note that there are two conformations for Glu 76. The distances (in Å) are shown.

and Schiff-base nitrogen (Fig. 2c and Supplementary Fig. 4), favouring the expected geometry for decarboxylation. This role for the carboxylate group in substrate alignment is consistent with the observed 250-fold decrease in k_{cat} (with no significant change in K_m) measured for the Glu76Gln mutant. Notably, the lack of a change in pH optimum for this mutant enzyme argues against an effect of Glu 76 on the ionization of the nucleophilic Lys (Supplementary Fig. 3).

In addition, in the unliganded and liganded complexes the side chain of Glu 76 has two alternative conformations (Fig. 2c). In one conformation Glu 76 (20% occupancy in the liganded enzyme subunit B and 100% occupancy in all subunits of the unliganded enzyme) is 4.0 Å from the β-carbonyl group of the inhibitor, whereas in the second it is 3.3 Å from the carboxylate of Glu 61 (positioned closer to the opening of the solvent channel). By analogy, in the substrate complex the side chain of Glu 76 may alternate between two positions, one proximal to Glu 61 (where it would be destabilized by like-charge repulsion) and one near to the β-carboxylate of the substrate Schiff base (favouring decarboxylation; Fig. 3). This model is supported by

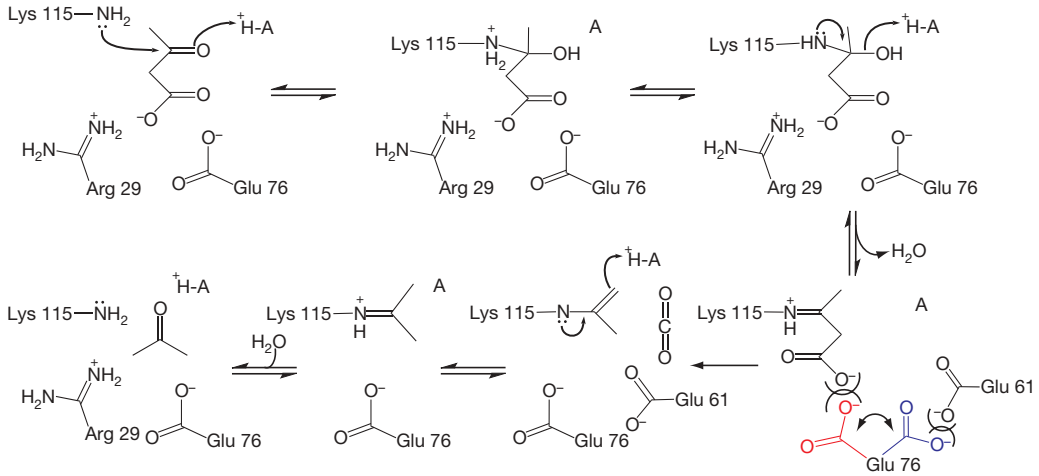


Figure 3 | The proposed mechanism of AADase. The two alternative conformations of the side chain of Glu 76 are shown in blue and red. A, acid.

the catalytic activity of the Glu61Gln mutant, which shows a decrease in k_{cat} (~20-fold with no change in K_m) similar in magnitude to that caused by the Glu76Gln mutation. The slight downward shift of the pH optimum of the Glu61Gln mutant (Supplementary Fig. 3) is not in the direction expected if this residue had an effect on the pK_a of Lys 115.

Oligomeric structure

The AADase homododecamer known to exist in solution[19] was generated by crystallographic symmetry from the protomers in the asymmetric unit; the resultant assembly did not differ between the structures from the two clostridial species. The 12 monomers form a tetrahedron, with axial lengths of 78 Å and longest dimension of 118 Å, enclosing a central tetrahedral cavity with axial lengths of 31 Å (Fig. 4a). The crystallographic dodecamer is consistent with dimensions determined for CaAAD by means of small angle X-ray scattering of 117 ± 3 Å in maximal diameter (Supplementary Figs 5 and 6). Additionally, two-dimensional cryo-electron microscopy projections of single CaAAD particles yielded a three-dimensional

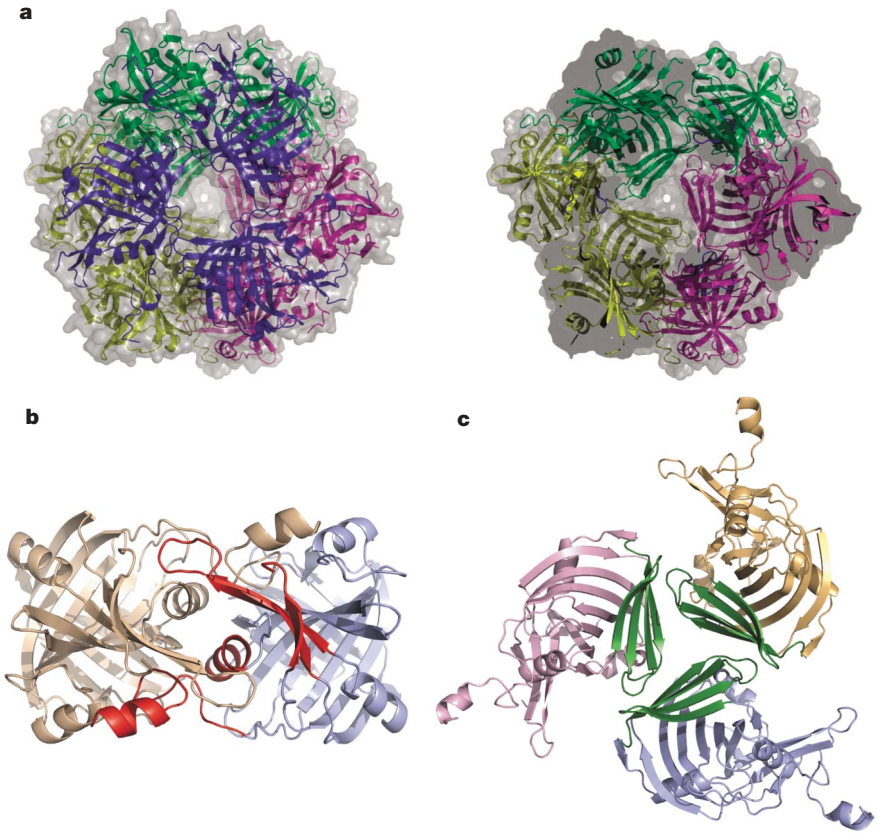


Figure 4 | The biological dodecamer of AADase built from the crystallographic asymmetric unit. a, Shown is the view parallel to the three-fold of the tetrahedral macromolecular assembly, with each trimer coloured differently (left) and cutaway to show the hollow interior (right). **b**, View of the dimer, with each subunit coloured separately and the helix and three-stranded sheet comprising the interface from a single subunit depicted in red. **c**, The structure is shown as a trimer, with each subunit coloured separately and the four-stranded sheet comprising the interface depicted in green (also shown in green in Fig. 1b).

volume with inner cavity dimensions (~30 Å) that did not differ from those determined from the crystallographic dodecamer (Supplementary Fig. 7). Thus, the crystallographic model appears to be an accurate representation of the state of the enzyme in solution. Macromolecular assemblies with tetrahedral geometry are extremely rare. The basic symmetry of a tetrahedron is that of four trimers, with each three-fold symmetry axis at a vertex, or that of six dimers with each two-fold at the centre of the axes connecting the vertices. The dimer interface comprising one three-stranded β-sheet and the amino-terminal α-helix of each monomer lies at the bottom of the β-cone and interdigitates with the identical interface of the neighbouring subunit (Fig. 4b). The trimer interface is formed from the four-stranded β-sheets, which form a platform with intersubunit interactions at the three-fold symmetry axis (Fig. 4c) by means of hydrophobic (CaAAD) or ionic (CvAAD) interactions.

The dimer interface forms extensive hydrophobic, hydrogen-bonding and ionic interactions involving more than 10% of the amino-acid residues of AADase (Supplementary Table 4 and Supplementary Fig. 8). It is these interactions that establish the dimeric form of CaAAD even in the presence of 4 M urea[19]. The conversion of acetoacetate to acetone is a key process in the anaerobic metabolism of carbohydrate in solventogenic bacteria such as *C. acetobutylicum*. The oligomeric state of AADase may be used to stabilize the enzyme at the very low pH and high solvent environment found in the cell during solventogenesis, or to colocalize the enzymes in the solventogenic pathway to facilitate the diffusion of product to the next enzyme in the pathway. The high K_m values for acetyl CoA acetyl transferase and AADase in the solventogenesis pathway (0.27 and 8 mM, respectively)[10,20] support this concept.

METHODS SUMMARY

The structure of AADase was determined by X-ray crystallography using phases determined experimentally by means of single wavelength anomalous diffraction on the selenomethionine-substituted protein. The dodecameric state of AADase greatly increased the complexity of *de novo* structure determination of the enzyme from *C. acetobutylicum* owing to the presence of near perfect merohedral twinning in many crystals and the presence of two dodecamers in the asymmetric unit (216 Se sites) in other crystal forms. The final successful structure analysis depended on determination of the structure of AADase from another bacterial source, CvAAD (Supplementary Table 2), which crystallizes with a smaller number of protomers in the asymmetric unit, and the use of the resulting model to determine the phases by means of molecular replacement of the native CaAAD and liganded CvAAD structures (see Supplementary Table 1 and Methods). Steady-state-kinetic analyses were performed on wild-type and mutant enzymes using a simple spectrophotometric assay. The dimensions of the oligomer of CaAAD in solution were determined by means of small angle X-ray scattering and further verified using three-dimensional particle reconstruction of cryo-electron microscopy images.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 September 2008; accepted 25 February 2009.

1. Westheimer, F. H. Coincidences, decarboxylation and electrostatic effects. *Tetrahedron* **51,** 3–20 (1995).
2. Hamilton, G. A. & Westheimer, F. H. On the mechanism of the enzymatic decarboxylation of acetoacetate. *J. Am. Chem. Soc.* **81,** 6332–6333 (1959).
3. Fridovich, I. & Westheimer, F. H. On the mechanism of the enzymatic decarboxylation of acetoacetate. II. *J. Am. Chem. Soc.* **84,** 3208–3209 (1962).
4. Laursen, R. A. & Westheimer, F. H. The active site of acetoacetate decarboxylase. *J. Am. Chem. Soc.* **88,** 3426–3430 (1966).
5. Warren, S., Zerner, B. & Westheimer, F. H. Acetoacetate decarboxylase. Identification of lysine at the active site. *Biochemistry* **5,** 817–823 (1966).
6. Kokesh, F. C. & Westheimer, F. H. A reporter group at the active site of acetoacetate decarboxylase. II. Ionization constant of the amino group. *J. Am. Chem. Soc.* **93,** 7270–7274 (1971).
7. Harris, T. K. & Turner, G. J. Structural basis of perturbed pK_a values of catalytic groups in enzyme active sites. *IUBMB Life* **53,** 85–98 (2002).
8. Gerlt, J. A. Obituary, Frank H. Westheimer. *Nature* **447,** 543 (2007).
9. Gallivan, J. P. & Dougherty, D. A. Cation–π interactions in structural biology. *Proc. Natl Acad. Sci. USA* **96,** 9459–9464 (1999).
10. Highbarger, L. A., Gerlt, J. A. & Kenyon, G. L. Mechanism of the reaction catalyzed by acetoacetate decarboxylase. Importance of lysine 116 in determining the pK_a of active-site lysine 115. *Biochemistry* **35,** 41–46 (1996).
11. García-Moreno, E. B. *et al.* Experimental measurement of the effective dielectric in the hydrophobic core of a protein. *Biophys. Chem.* **64,** 211–224 (1997).
12. Lee, J. K. & Houk, K. N. A proficient enzyme revisited: the predicted decarboxylase mechanism for orotidine monophosphate. *Science* **276,** 942–945 (1997).
13. Rashin, V. & Honig, B. Reevaluation of the Born model of ion hydration. *J. Phys. Chem.* **89,** 5588–5593 (1985).
14. Stites, W. E., Gittis, A. G., Lattman, E. E. & Shortle, D. In a staphylococcal nuclease mutant the side-chain of a lysine replacing valine66 is fully buried in the hydrophobic core. *J. Mol. Biol.* **221,** 7–14 (1991).
15. Barbas, C. F. III *et al.* Immune versus natural selection: antibody aldolases with enzymic rates but broader scope. *Science* **278,** 2085–2092 (1997).
16. Crosby, J., Stone, R. & Liehard, G. E. Mechanisms of thiamine-catalyzed reactions. Decarboxylation of 2-(1-carboxy-1-hydroxyethyl)-3,4-dimethylthiazolium chloride. *J. Am. Chem. Soc.* **92,** 2891–2900 (1970).
17. Fridovich, I. A study of the interaction of acetoacetic decarboxylase with several inhibitors. *J. Biol. Chem.* **243,** 1043–1051 (1968).
18. O'Leary, M. H. in *Mechanisms of Catalysis* (ed. Sigman, D. S.) 239 (Academic, 1992).
19. Tagaki, W. & Westheimer, F. H. Acetoacetate decarboxylase. Reassociation of subunits. *Biochemistry* **7,** 891–894 (1968).
20. Wiesenborn, D. P., Rudolph, F. B. & Papoutsakis, E. T. Thiolase from *Clostridium acetobutylicum* ATCC 824 and its role in the synthesis of acids and solvents. *Appl. Environ. Microbiol.* **54,** 2717–2722 (1988).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Murzin for provisional classification of the AADase fold and C. Akey for valuable advice on the execution and interpretation of the electron microscopy work. We also thank H. Robinson and A. Soares for help with data collection. This work was supported by a grant to K.N.A. from the National Science Foundation. Data for this study were measured at Beamlines X12B, X25 and X29A of the National Synchrotron Light Source. Financial support comes principally from the Offices of Biological and Environmental Research (BER) and of Basic Energy Sciences (BES) of the US Department of Energy (DOE), and from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH). Small-angle X-ray scattering analyses were carried out at the Stanford Synchrotron Radiation Lightsource, funded by DOE, BES. The SSRL Structural Molecular Biology Program is supported by DOE, BER, and by NIH, NCRR. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH.

Author Contributions M.-C.H. cloned, expressed, purified, crystallized, collected data and performed crystal structure determination, refinement and model analysis. H.T. designed, executed, analysed and wrote the description of the small-angle X-ray scattering analysis. J.F.M. designed, executed, analysed and wrote the description of the electron microscopy experiments. K.N.A. conceived of and designed the project. M.-C.H. and K.N.A. wrote the manuscript; all authors discussed the results and commented on the manuscript.

Author Information The coordinates and structure factors have been deposited in the Protein Data Bank with accession codes 3BH2, 3BGT and 3BH3 corresponding, respectively, to *C. acetobutylicum* acetoacetate decarboxylase and *C. violaceum* acetoacetate decarboxylase in the unliganded form and complexed with 2,4-pentanedione. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.N.A. (drkallen@bu.edu).

METHODS

Materials. Except where indicated, all chemicals were purchased from Sigma-Aldrich or American Bioanalytical. DNase I was from Worthington Biochemical Corp. All enzymes and primers for PCR, T4 DNA ligase and restriction enzymes were from Invitrogen. The DNA cleanup kit and miniprep kit were from Qiagen. Host cells and pET vectors were purchased from Novagen.

Cloning, expression and purification. The plasmid encoding the AADase gene of *C. acetobutylicum* (CaAAD) was purified from host cells[10] provided by J. Gerlt. Primers (5′-GCGCCCATGGTAAAGGATGAAGTAAT-3′ and 5′-CGCGGATCCTTACTTAAGATAATCATATAT-3′) containing NCOI and BamHI endonuclease cleavage sites (underlined) were used to amplify the gene by PCR. The genome of *C. violaceum* was purchased from American Type Culture Collection (ATCC 12472D) and treated with the endonuclease BamHI before PCR. The AADase from *C. violaceum* (CvAAD) gene was amplified using the primers (5′-TGAAGGGAGTTACATATGAAGCAACAGGAAGTCCAG-3′ and 5′-GGGGCCGTTTTTCTCGGATCC-GCCGTTATTGCCGCAGG-3′) containing NdeI and BamHI restriction sites (underlined). The CaAAD and CvAAD were cloned into pET3d and pET15b vectors, respectively. The ligation products were transformed into *Escherichia coli* DH5α and sequenced by the Tufts University Core Facility.

CaAAD was expressed and purified from *E. coli* BL21 (DE3) cells using a modification of the procedure described previously[10]. The transformed cells were grown in Luria Bertani medium at 37 °C and induced by 1 mM isopropyl-β-D-thiogalactopyranoside for 20 h at room temperature (25 °C). The cells were collected by centrifugation. The cell pellet was suspended in 40 ml ice-cold lysis buffer (50 mM phosphate buffer, pH 7.4, 2 mM DTT, 1 mM PMSF and 1 mg DNase I) and lysed by sonication (5 × 6 min, 80% power and 60% duty cycle, Branson Sonifier 250). The cell lysate were centrifuged at 4 °C for 1 h at 147,000*g*. The protein was precipitated by dropping the pH to ~3.8–3.9, and the precipitated material was dissolved in 150 ml 200 mM phosphate buffer, pH 5.95, at 4 °C overnight (12 h). The remaining impurities were removed by precipitating with 55% $(NH_4)_2SO_4$ followed by centrifugation at 3,700*g* for 30 min. CaAAD was precipitated with 75% $(NH_4)_2SO_4$, dissolved in 3 ml of 50 mM phosphate buffer, pH 5.95, and loaded directly onto a S-200 gel filtration column (GE Healthcare) pre-equilibrated with buffer A (50 mM phosphate buffer, pH 5.95). The fractions containing CaAAD, detected using a kinetic assay (see below) and 12% SDS–PAGE, were collected. If the CaAAD fractions from S-200 gel filtration were >90% pure, the next purification step, using DEAE-Sepharose, was omitted. If not, the collected fractions were loaded onto a DEAE-Sepharose column (GE Healthcare) equilibrated with buffer A. The column was washed with buffer A and eluted with a 0.5 l linear gradient of 0–0.5 M NaCl in buffer A. All fractions containing CaAAD, detected by activity assays and 12% SDS–PAGE, were collected and dialysed against 5 mM phosphate buffer, pH 5.95, and 2 mM DTT at 4 °C overnight, followed by concentration to ~10 mg ml^{-1} using an Amicon ultracentrifugal filter device (Millipore).

The expression and purification of CvAAD was similar to CaAAD with some modification. Forty minutes before induction, 1× Augmedium (AthenaES) was added to reduce the percentage of protein expressed as insoluble inclusion bodies. The cells were collected by centrifugation and lysed by sonication in 50 ml ice-cold lysis buffer (50 mM HEPES, pH 7.4, 30 mM NaCl, 1 mM PMSF and 1 mg DNase I). The cell lysate was centrifuged at 4 °C for 1 h at 147,000*g* and the supernatant loaded onto a column packed with 15 ml TALON cobalt resin (Clontech). The column was washed with 50 ml wash buffer (50 mM HEPES, pH 7.4, 30 mM NaCl and 2 mM β-mercaptoethanol) and 50 ml wash buffer plus 30 mM imidazole, pH 7.6, and eluted with 50 ml elution buffer (50 mM HEPES pH 7.4, 30 mM NaCl, 150 mM imidazole pH 7.6 and 2 mM β-mercaptoethanol). The protein was dialysed against 5 mM imidazole buffer, pH 7.6, and 2 mM DTT at 4 °C overnight and concentrated to ~20 mg ml^{-1} using an Amicon Ultra centrifugal filter device.

Selenomethionine-labelled CvAAD (SeMet CvAAD) was expressed in *E. coli* (B834) cells. The cells were grown in SelenoMet medium base (AthenaES). A custom Augmedium formulation (0.5 g l^{-1} methionine was replaced by 0.15 g l^{-1} selenomethionine) was added 2 h before induction. The cells were induced with 1 mM isopropyl-β-D-thiogalactopyranoside at 18 °C for 48 h and collected by centrifugation. The purification of selenomethionine-labelled CvAAD was the same as that of native CvAAD.

Mutageneses of CvAAD was accomplished by means of quick-change site-directed mutagenesis (Invitrogen) using CvAAD/pET3A as the template. The R29Q mutant was created using the primers (5′-GCCTTATCGTTTCGTCAACCAGGAATACATGATCATCACCTA-3′ and 5′-TAGGTGATGATCATGTATTCCTGGTTGACGAAACGATAAGGC-3′). The E61Q mutant was created using the primers (5′-GGATTTGGCGATTACTCGCAAAGCGGGCAGG-3′ and 5′-CCTAAACCGCTAATGAGCGTTTCGCCCGTCC-3′). The E76Q mutant was created using the primers (5′-GCCTTATCGTTTCGTCAACCAGGAATACATGATCATCACCTA-3′, and 5′-TAGGTGATGATCATGTATTCCTGGTTGACGAAACGATAAGGC-3′). The mutated sites are underlined. The resulting PCR products were digested with NdeI and BamHI endonuclease, and cloned into pET15b vectors that were cut by the same enzymes. The ligation products were transformed into *E. coli* DH5α, followed by DNA sequencing by Agencourt Bioscience. The plasmid, encoding His_6-tagged CvAAD point mutants, were purified from *E. coli* DH5α and transformed into *E. coli* BL21(DE3) for protein expression. The expression and purification of CvAAD mutants was identical to that of wild-type CvAAD.

Crystallization and data collection. The initial crystallization conditions of CaAAD and CvAAD were identified by high-throughput screening (Hauptman Woodward Medical Research Institute) and refined by the hanging-drop vapour-diffusion method. The final crystallization condition of CaAAD was 18–20% glycerol, 40 mM phosphate buffer, pH 5.95, 100 mM sarcosine and 14–15% PEG 3350 (Hampton Research). Drops were formed by mixing 1–2 μl of protein solution (10 mg ml^{-1}) with 1 μl of well solution. Crystals of CaAAD were flash-frozen in a gaseous N_2 stream at 100 K. X-ray diffraction data were collected at the National Synchrotron Light Source (Brookhaven National Laboratory, New York), Beamline X25. CaAAD crystallized in space group *R*32 with unit-cell dimensions $a = b = 104.07$ Å, $c = 578.09$ Å and diffracted to 2.4 Å resolution.

Crystals of CvAAD and SeMet CvAAD were grown by hanging-drop vapour diffusion over a well solution containing 0.4–0.5 M K_2HPO_4/NaH_2PO_4 at pH 7.5–7.9. Drops consisted of 1–2 μl of protein solution (8–12 mg ml^{-1}) and 1 μl of well solution. Crystals were transferred to well solution with 30% glucose as a cryo-protectant and crystals were frozen in liquid N_2. The automounter at BNL, Beamline X12B, was used to screen crystals for diffraction and data were collected at Beamline X29A. SeMet CvAAD crystallized in space group *R*3 ($a = b = 105.45$ Å, $c = 252.38$ Å). A 2.1 Å multiple wavelength anomalous diffraction (MAD) data set was collected (0.9791 Å, 0.9793 Å and 0.9754 Å), but only the data collected at 0.9791 Å were used to solve the phases. For the 2,4-pentanedione complex structure with CvAAD, the enzyme (10 mg ml^{-1}) was co-crystallized with 2 mM potassium acetyl acetonate under the same condition as unliganded CvAAD. The crystal was additionally soaked in mother liquid plus 120 mM potassium acetyl acetonate for 1 h at room temperature and transferred to well solution with 30% glucose plus 4 mM potassium acetyl acetonate as cyro-protectant and frozen in liquid N_2. The inhibitor-soaked crystals diffracted to 2.1 Å and were isomorphous to the SeMet CvAAD. A 2.1 Å data set was collected at 0.900 Å at Brookhaven National Laboratories, beamline X12B. All data sets were indexed and scaled using HKL2000 (ref. 21). Data collection statistics are summarized in Supplementary Table 1.

Structure determination and model refinement. The structure of CvAAD was phased by the single wavelength anomalous diffraction method, using the 2.1 Å SeMet CvAAD data set collected at the peak wavelength 0.9791 Å. The Se substructure was solved using HKL2MAP/SHELXD[22,23], which identified 24 of 32 selenomethionine sites. The 24 sites were used to phase the protein structure in SOLVE[24] and generated a high-quality electron density map. The map was improved using RESOLVE[25], by means of NCS averaging (4 monomers per asymmetric unit). An initial model including 44% of the structure was built automatically by RESOLVE. The experimentally phased NCS averaged electron density map permitted further manual rebuilding of one of the four monomers using COOT[26]. This monomer was then used to perform molecular replacement using MOLREP[27] to find the remaining three monomers in the asymmetric unit. Iterative rounds of model building and refinement were performed using COOT, CNS and REFMAC5[26–28].

The structure of CaAAD was phased by molecular replacement with the program PHASER using SeMet CvAAD as the search model[29]. The process of refinement and rebuilding was the same as that used for SeMet CvAAD. The structure of CaAAD and 2,4 pentanedione liganded CvAAD was phased by molecular replacement with the program PHASER[29] using the structure of SeMet CvAAD as the search model. The process of refinement and rebuilding was the same as that used for SeMet CvAAD. The model of the inhibitor was not added to the refinement until the value of R_{work} dropped below 28% to avoid phase bias. The topology restraints for 2,4-pentanedione adduct of Lys 115 for use in REFMAC5 were generated using SKETCHER in the CCP4 program suite[30]. The Ramachandran plot, as assessed by Procheck[31], had 99.5% of the residues of unliganded CaAAD and CvAAD and 99.2% of the residues of liganded CvAAD in the allowed regions with 1 residue (Ser 126 in CaAAD/Gln 127 in CvAAD) in the disallowed regions for all three structures.

Small-angle X-ray scattering measurements. Solution X-ray scattering measurements were performed at the Stanford Synchrotron Radiation Lightsource Beamline 4-2 (ref. 32). Each protein sample solution containing 50 mM phosphate (pH 5.95) and 2 mM DTT was held in a sample cuvette, maintained at 20 °C and located at 2.5 m from a MarCCD165 detector

(MarUSA). The detector pixel numbers were converted to the momentum transfer $Q = 4\pi\sin(\theta)/\lambda$, where θ is one half of the scattering angle and λ the X-ray wavelength 1.381 Å, using the (100) reflection and related reflections recorded from a cholesterol myristate powder sample. Twenty-four 10-s exposures were acquired in series at 2, 5, 7.5, 10 and 15 mg ml^{-1} (Supplementary Fig. 5). No time-dependent radiation-induced protein aggregation was observed and all 24 images were averaged after intensity scaling, with the exception of 10 and 15 mg ml^{-1} samples, for which the last few exposures were excluded in the averaging process due to the onset of radiation-induced aggregation. The radial integration, intensity-scaling, statistical analysis, frame-averaging and background subtraction were done by MarParse[32]. Radii of gyration (R_g) and forward scattered intensities ($I_{(Q=0)}$) were obtained by Primus[33] in the Q range 0.0117–0.0297 Å^{-1} (Supplementary Fig. 5, inset). The concentration-scaled forward scattered intensity remained within 3% of the mean value between 2 and 10 mg ml^{-1}, indicating no evidence of concentration-dependent aggregation. Slightly lower R_g values at higher protein concentrations indicated mildly repulsive inter-particle interaction. The composite scattering curve with satisfactorily high statistics covering Q values 0.0117 to 0.25 Å^{-1} was thus obtained by scaling and merging the scattering curve recorded at 2 mg ml^{-1} ($0.0117 < Q < 0.0611$ Å^{-1}) with the intermediary angle data recorded at 15 mg ml^{-1} ($0.0448 < Q < 0.25$ Å^{-1}) by Primus[33] to minimize weak effects of the particle structure factor due to the repulsive interaction (Supplementary Fig. 6). The overlapping Q range contained 30 common data points to assure accuracy of the scaling. The X-ray scattering profile and the electron pair distance distribution function $P(r)$ were computed for the crystallographic structure model using ORNL_SAS[34]. The electron density and thickness of the hydration layer were altered in the range 3–15% higher electron density of water and 2.5–3.5 Å. The best fit was obtained with 8% excess density and 2.5 Å layer thickness, although similarly satisfactory fits could be obtained with lower hydration layer density and a slightly thicker hydration layer. Q values in the experimental data were also adjusted by 1–3% to account for the small unit cell parameter variations observed in the crystallographic study. The lowest chi square value we have obtained with 1% Q value adjustment was 2.445, which confirms the presence of the crystallographic dodecamer in solution but reflects the clearly recognizable deviation in the Q range 0.06–0.10 Å^{-1}. This deviation may suggest minor structural differences between solution and crystal structures at the tertiary and/or quaternary structural level because experimental conditions can not be identical. Similar differences between experimental and computed X-ray scattering curves have been observed for other oligomeric enzyme systems[35]. The indirect Fourier transform analysis of the composite scattering curve by GNOM[36] yielded the electron pair distance distribution function, giving the maximum particle dimension of 117 (± 3 (± s.d.)) Å and R_g of 42.9 Å, both of which are highly consistent with the corresponding values of 118 and 43.3 Å obtained for the crystal structure with the hydration layer (Supplementary Fig. 6). The nearly perfect match of the experimental and computed $P(r)$ indicates that the solution and crystal structures are essentially identical at the quaternary structure level (Supplementary Fig. 6, inset).

Cryo-electron microscopy and image processing. The cryo-electron microscopy and image processing of CaAAD were performed at Boston University School of Medicine. CaAAD (10 mg ml^{-1}, 1 mM phosphate buffer, pH 5.95) was loaded onto perforated carbon film on 400 mesh copper grids. The specimens were plunge-frozen in liquid ethane cooled by liquid nitrogen[37] and data collected on a Tecnai TF20 with Kodak SO163 film at Boston University School of Medicine. Electron micrographs were recorded at 200 kV at ×62,000 with a defocus range of −1.0 to −2.5 μm. Negatives were digitized on a CreoScitex EVERSMART scanner using a 4.54 μm raster and binned 4 × 4 to 2.93 Å per pixel. Particles were picked with boxer in EMAN[38] and classified using refine2d in EMAN to avoid bias from a three-dimensional model[39].

Additionally, the three-dimensional structure was refined in EMAN using ~8,700 particles and by imposing the tetrahedral symmetry found in the crystal structure[39]. The resolution was estimated by splitting the data set into two halves (according to even and odd numbered particles), generating an even and an odd volume and then calculating their cross-correlation coefficient. Plotting the correlation coefficient versus resolution gave the Fourier shell correlation (FSC) curve. Using the FSC 0.5 criteria, it was concluded that the resolution of the three-dimensional volume was 12 Å.

Steady-state kinetics. The decarboxylation of acetoacetate by CvAAD and CaAAD was assayed by monitoring the disappearance of the enolate form of acetoacetate spectrophotometrically at 270 nm ($\Delta\varepsilon = 26.7\,M^{-1}\,cm^{-1}$) at 25 °C in 50 mM phosphate buffer, pH 5.95, on a Beckman Coulter DU-800 spectrophotometer[10]. The same assay was used for CvAAD mutants in 50 mM phosphate/citrate buffer at various pH values on a Varian Carey-300 spectrophotometer. Data were fitted to the Michaelis–Mention equation $V = V_{max}[S]/(K_m + [S])$, where V is the initial rate at substrate concentration S, and K_m is the Michaelis constant. k_{cat} is obtained from V_{max} and total enzyme concentration in the reaction. The k_{app} for wild-type and mutant CvAAD proteins versus pH was obtained from the initial rate and enzyme concentration at the saturating substrate concentration of 20 mM, with the exception of the CvAAD R29Q mutant for which activity could only be measured using 40 mM substrate.

21. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
22. Pape, T. & Schneider, T. R. HKL2MAP: a graphical user interface for phasing with SHELX programs. *J. Appl. Crystallogr.* **37,** 843–844 (2004).
23. Usón, I. & Sheldrick, G. M. Advances in direct methods for protein crystallography. *Curr. Opin. Struct. Biol.* **9,** 643–648 (1999).
24. Terwilliger T.C. & Berendzen J.. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55,** 849–861 (1999).
25. Terwilliger, T. C. Maximum likelihood density modification. *Acta Crystallogr. D* **56,** 965–972 (2000).
26. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphic. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
27. Collaborative computation project number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
28. Brünger, A. T. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
29. Storoni, L. C., McCoy, A. J. & Read, R. J. Likelihood-enhanced fast rotation function. *Acta Crystallogr. D* **60,** 432–438 (2004).
30. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59,** 1131–1137 (2003).
31. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. Procheck: a program to check the sterochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291 (1993).
32. Smolsky, I. L. *et al.* Biological small-angle x-ray scattering facility at the Stanford Synchrotron Radiation Laboratory. *J. Appl. Crystallogr.* **40,** S453 (2007).
33. Konarev, P. V. *et al.* PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36,** 1277–1282 (2003).
34. Tjioe, E. & Heller, W. T. *ORNL_SAS:* software for calculation of small-angle scattering intensities of proteins and protein complexes. *J. Appl. Crystallogr.* **40,** 782–785 (2007).
35. Svergun, D. I. *et al.* Large differences are observed between the crystal and solution quaternary structures of allosteric aspartate transcarbamylase in the R state. *Proteins* **27,** 110–117 (1997).
36. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25,** 495–503 (1992).
37. Ménétret, J. F. *et al.* The structure of ribosome–channel complexes engaged in protein translocation. *Mol. Cell* **6,** 1219–1232 (2000).
38. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128,** 82–97 (1999).
39. Ludtke, S. J., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure* **12,** 1129–1136 (2004).

Vol 459|21 May 2009|doi:10.1038/nature08022

LETTERS

Non-radial oscillation modes with long lifetimes in giant stars

Joris De Ridder[1], Caroline Barban[2], Frédéric Baudin[3], Fabien Carrier[1], Artie P. Hatzes[4], Saskia Hekker[5,1], Thomas Kallinger[6], Werner W. Weiss[6], Annie Baglin[2], Michel Auvergne[2], Réza Samadi[2], Pierre Barge[7] & Magali Deleuil[7]

Towards the end of their lives, stars like the Sun greatly expand to become red giant stars. Such evolved stars could provide stringent tests of stellar theory, as many uncertainties of the internal stellar structure accumulate with age. Important examples are convective overshooting and rotational mixing during the central hydrogen-burning phase, which determine the mass of the helium core, but which are not well understood[1]. In principle, analysis of radial and non-radial stellar oscillations can be used to constrain the mass of the helium core. Although all giants are expected to oscillate[2], it has hitherto been unclear whether non-radial modes are observable at all in red giants, or whether the oscillation modes have a short or a long mode lifetime[3–7], which determines the observational precision of the frequencies. Here we report the presence of radial and non-radial oscillations in more than 300 giant stars. For at least some of the giants, the mode lifetimes are of the order of a month. We observe giant stars with equally spaced frequency peaks in the Fourier spectrum of the time series, as well as giants for which the spectrum seems to be more complex. No satisfactory theoretical explanation currently exists for our observations.

Stochastic oscillations with small amplitudes have been firmly detected in a few bright red giants of spectral types G and K, with both spectroscopic and photometric data[3–16]. However, the exact information contained in the power spectra of red giants is still much debated. This is well illustrated by the star ε Oph, the only red giant up to now for which ground-based radial velocity data[3] as well as 28 days of contiguous space-based photometric data[4] have been gathered.

Two completely different ways of interpreting the power spectrum of this star have been presented. One interpretation[3,4] advocates the presence of a single comb of broad equidistant peaks of radial modes only. A direct fit with Lorentz profiles of these peaks leads to a mode lifetime of approximately 2.7 days. The second interpretation[5], however, advocates the presence of at least 21 independent narrow-lined modes, both radial and non-radial, with a lifetime between 10 and 20 days. Many of the peaks that in the former interpretation are considered to be part of the wings of a stochastic realization of a broad Lorentz profile, are thus considered independent oscillation modes in the second interpretation. Up to now, no consensus has been reached in the literature whether the mode lifetimes in ε Oph, or any other G or K giant, are short or long. We refer to the Supplementary Discussion for an outline of arguments in favour of either interpretation. From this overview it is clear that there is neither consensus nor a proper understanding of stochastic oscillations in red giants. Moreover, all present studies deal with a small number of (often the same) giants. This is because gathering time series of red giants of sufficiently high quality is challenging. A larger sample of high-quality time series of red giants is the only way to gain a better understanding.

Here we present such a sample of high-precision photometric time series measured by the satellite CoRoT[17]. The Supplementary Methods supplies details on this mission, and an outline of the data reduction steps that we performed. To identify the red giants among the observed targets, we devised a semi-automated classification algorithm that relies on the power spectrum of the targets. We first selected the targets brighter than Johnson V magnitude $m_V = 15$, because simulations showed that for fainter giants the signal-to-noise ratio would not allow us easily to detect a power excess. For a target to be an acceptable red giant candidate, we required first that its power spectrum should show a background noise with increasing amplitude at low frequencies, which is what we expect because of granulation, similar to that of the Sun[18]. In addition, we required that a single power excess due to oscillations must be present, with a position between 10 and 120 μHz (ref. 19) and a width of at least 5 μHz. An increasing noise level at low frequencies due to granulation is a necessary but not a sufficient condition, because instrumental noise may show the same signature. Finally, we always verified by eye that possible frequency peaks linked to the satellite's orbit (owing to stray-light, for example, or the satellite's passage through the South Atlantic Anomaly) did not affect the classification. Following the procedure outlined above, we retained more than 300 candidate red giant pulsators. The colour-magnitude diagram shown in Supplementary Fig. 1 confirms that these pulsators are indeed located on the red giant branch.

The selected red giants show a large variety of power spectra. This is demonstrated in Fig. 1 where a stack of Fourier power spectra of nine pulsating red giants is presented. The power spectra show the unprecedented low noise level at the higher frequencies where red giant oscillations can be easily detected and analysed. A particularly interesting power density spectrum is the one of the red giant CoRoT-101034881, presented in Fig. 2, which shows a regular pattern of oscillation peaks. Folding the power spectrum leads to the echelle diagram shown in Fig. 3 in which the 12 modes form three 'ridges' corresponding to the radial, dipole and quadrupole modes. The power spectrum of this giant therefore provides clear evidence for the existence of non-radial modes. Three more examples of such giants are presented in Supplementary Figs 2–4.

It should be noted, however, that the theoretical spectrum of non-radial modes in red giants is much denser than what we observe here[2,20]. Presumably only the non-radial modes that are standing waves in the outer oscillation cavity of the giant are visible, while

[1]Instituut voor Sterrenkunde, K. U. Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium. [2]LESIA, UMR8109, Université Pierre et Marie Curie, Université Denis Diderot, Observatoire de Paris, 92195 Meudon, France. [3]Institut d'Astrophysique Spatiale, UMR 8617 CNRS-Université Paris XI, Campus d'Orsay, F-91405 Orsay Cedex, France. [4]Thüringer Landessternwarte, D-07778 Tautenburg, Germany. [5]Royal Observatory of Belgium, Ringlaan 3, 1180 Brussels, Belgium. [6]Institute for Astronomy, University of Vienna, Türkenschanzstrasse 17, A-1180 Vienna, Austria. [7]Laboratoire d'Astrophysique de Marseille, OAMP, Université Aix-Marseille & CNRS, 38 rue Frédéric Joliot Curie, 13388 Marseille 13, France.

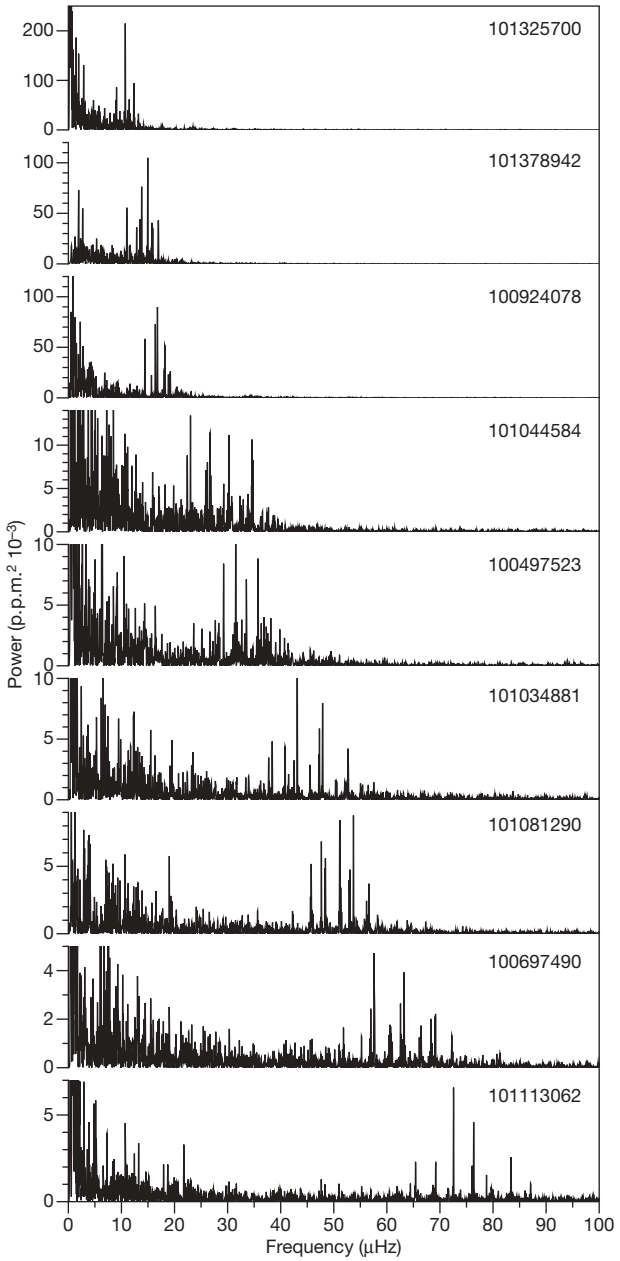


Figure 1 | A stack of power spectra of nine red giant pulsators. The power is expressed in parts per million (p.p.m.) squared divided by 1,000. The oscillation frequency peaks are clearly visible, around 75 μHz for the bottom panel, down to 10 μHz for the top panel. The values for the corresponding frequency ν_{max} of maximum oscillation power are consistent with what is expected from scaling laws[19]. At low frequencies (<10 μHz) our detection is limited by the granulation noise, and at high frequencies (>80 μHz) the amplitudes of the oscillations drop eventually under the threshold of the detection algorithm. The nine-digit numbers given are the CoRoT identifiers of the targets. We refer to Supplementary Table 1 for their USNO-A2 identifiers.

the non-radial modes oscillating in the inner cavity, which would make up a 'forest' of frequency peaks, are not visible. Another peculiarity of the power density spectrum shown in Fig. 2 is the narrowness of the frequency peak profiles. In fact, fitting the modes with Lorentz profiles yields widths that are close to the widths expected for a finite time series of about 150 days, indicating a mode lifetime of at least 50 days. This result contrasts sharply with some of the interpretations of observational results found for the giant ξ Hya (ref. 7) and, to some extent, for ε Oph (refs 3, 4).

Not all red giants observed by CoRoT show a power density spectrum as clear as that of CoRoT-101034881. As a contrasting example

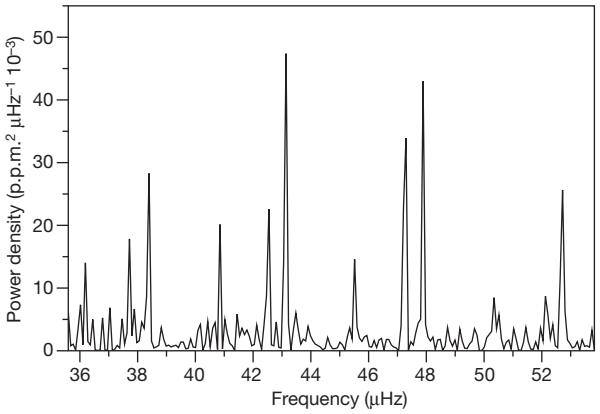


Figure 2 | Power density spectrum of the red giant candidate CoRoT-101034881 showing a frequency pattern with a regular spacing. This spacing is predicted by the theoretical asymptotic relation for high-order and low-degree oscillations[21]. Using the auto-correlation function of the power spectrum, we derive the large separation to be 4.8 μHz. This value is consistent with what is expected for red giants from scaling laws[19].

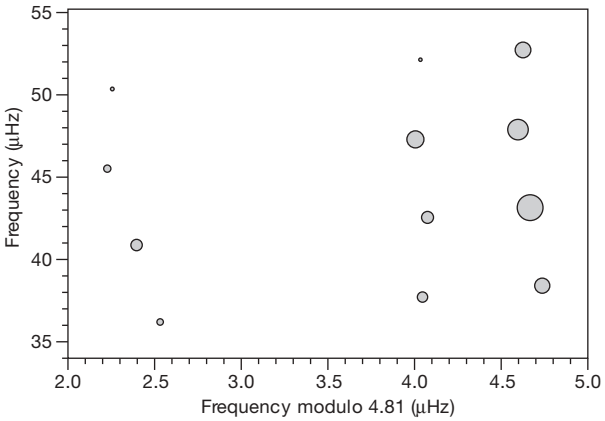


Figure 3 | Echelle diagram of the modes shown in Fig. 2, showing 'ridges' related to radial and non-radial modes. The folding frequency is 4.81 μHz. The size of the symbols is proportional to the height of the peak in the spectrum shown in Fig. 2. From the theoretical asymptotic relation for high-order and low-degree oscillations[21], we conclude that the three vertical ridges correspond to dipole modes (left), quadrupole modes (middle) and radial modes (right).

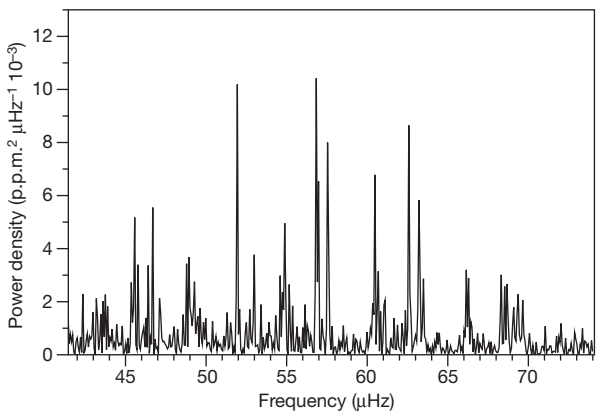


Figure 4 | Power density spectrum of the red giant CoRoT-101600807 showing broad profiles. This spectrum contrasts with that of CoRoT-101034881, highlighting the variety of red giant power spectra observed with CoRoT.

we show the power spectrum of the candidate red giant CoRoT-101600807 in Fig. 4, which also contains broad features. If these features correspond to single modes, this would imply mode lifetimes that are considerably shorter than those of CoRoT-101034881. However, if they correspond to many modes with narrow profiles, then this may imply that at least some of the non-radial modes trapped in the core of the giant are excited to detectable surface amplitudes.

Received 25 November 2008; accepted 26 March 2009.

1. Aerts, C., Christensen-Dalsgaard, J., Cunha, M. & Kurtz, D. W. The current status of asteroseismology. *Sol. Phys.* **251,** 3–20 (2008).
2. Dziembowski, W., Gough, D. O., Houdek, G. & Sienkiewicz, R. Oscillations of α UMa and other red giants. *Mon. Not. R. Astron. Soc.* **382,** 601–610 (2001).
3. De Ridder, J. *et al.* Discovery of solar-like oscillations in the red giant ε Ophiuchi. *Astron. Astrophys.* **448,** 689–695 (2006).
4. Barban, C. *et al.* Detection of solar-like oscillations in the red giant star ε Ophiuchi by MOST spacebased photometry. *Astron. Astrophys.* **468,** 1033–1038 (2007).
5. Kallinger, T. *et al.* Nonradial p-modes in the G9.5 giant ε Ophiuchi? Pulsation model fits to MOST photometry. *Astron. Astrophys.* **478,** 497–505 (2008).
6. Hekker, S., Aerts, C., De Ridder, J. & Carrier, F. Pulsations detected in the line profile variations of red giants. Modelling of line moments, line bisector and line shape. *Astron. Astrophys.* **458,** 931–940 (2006).
7. Stello, D., Kjeldsen, H., Bedding, T. R. & Buzasi, D. Oscillation mode lifetimes in ξ Hydrae: will strong mode damping limit asteroseismology of red giant stars? *Astron. Astrophys.* **448,** 709–715 (2006).
8. Edmonds, P. D. & Gilliland, R. L. K giants in 47 Tucanae: Detection of a new class of variable stars. *Astrophys. J.* **464,** L157–L160 (1996).
9. Buzasi, D. *et al.* The detection of multimodal oscillations on α Ursae Majoris. *Astrophys. J.* **532,** L133–L136 (2000).
10. Frandsen, S. *et al.* Detection of solar-like oscillations in the G7 giant star xi Hya. *Astron. Astrophys.* **394,** L5–L8 (2002).
11. Retter, A. *et al.* Oscillations in Arcturus from WIRE photometry. *Astrophys. J.* **591,** L151–L154 (2003).
12. Tarrant, N. J. *et al.* Asteroseismology of red giants: photometric observations of Arcturus by SMEI. *Mon. Not. R. Astron. Soc. Lett.* **382,** L48–L52 (2007).
13. Stello, D. *et al.* Multisite campaign on the open cluster M67—II. Evidence for solar-like oscillations in red giant stars. *Mon. Not. R. Astron. Soc.* **377,** 584–594 (2007).
14. Hatzes, A. P. & Zechmeister, M. The discovery of stellar oscillations in the planet-hosting giant star β Geminorum. *Astrophys. J.* **670,** L37–L40 (2007).
15. Stello, D., Bruntt, H., Preston, H. & Buzasi, D. Oscillating K giants with the WIRE satellite: determination of their asteroseismic masses. *Astrophys. J.* **674,** L53–L56 (2008).
16. Gilliland, R. L. Photometric oscillations of low-luminosity red giant stars. *Astron. J.* **136,** 566–579 (2008).
17. Baglin, A. *et al.* The CoRoT mission and its scientific objectives. *AIP Conf. Proc.* **895,** 201–209 (2007).
18. Pallé, P. L. *et al.* A measurement of the background solar velocity spectrum. *Astrophys. J.* **441,** 952–959 (1995).
19. Kjeldsen, H. & Bedding, T. R. Amplitudes of stellar oscillations: the implications for asteroseismology. *Astron. Astrophys.* **293,** 87–106 (1995).
20. Guenther, D. *et al.* Evolutionary model and oscillation frequencies for α Ursae Majoris: a comparison with observations. *Astrophys. J.* **530,** L45–L48 (2000).
21. Tassoul, M. Asymptotic approximations for stellar nonradial pulsations. *Astrophys. J. Suppl. Ser.* **43,** 469–490 (1980).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J.D.R., F.C. and S.H. acknowledge support from the Fund for Scientific Research, Flanders, from the research council of K. U. Leuven, and from the Belgian Federal Science Policy. A.P.H. acknowledges the support of the Deutsches Zentrum für Luft- und Raumfahrt. T.K. and W.W.W. acknowledge support by the Austrian Research Promotion Agency (FFG-ARL). J.D.R. thanks A. Miglio, M.-A. Dupret and C. Aerts for discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.D.R. (joris@ster.kuleuven.ac.be).

LETTERS

Stability against freezing of aqueous solutions on early Mars

Alberto G. Fairén[1], Alfonso F. Davila[1], Luis Gago-Duport[2], Ricardo Amils[3,4] & Christopher P. McKay[1]

Many features of the Martian landscape are thought to have been formed by liquid water flow[1,2] and water-related mineralogies on the surface of Mars are widespread and abundant[3]. Several lines of evidence, however, suggest that Mars has been cold with mean global temperatures well below the freezing point of pure water[4]. Martian climate modellers[5,6] considering a combination of greenhouse gases at a range of partial pressures find it challenging to simulate global mean Martian surface temperatures above 273 K, and local thermal sources[7,8] cannot account for the widespread distribution of hydrated and evaporitic minerals throughout the Martian landscape[3]. Solutes could depress the melting point of water[9,10] in a frozen Martian environment, providing a plausible solution to the early Mars climate paradox. Here we model the freezing and evaporation processes of Martian fluids with a composition resulting from the weathering of basalts, as reflected in the chemical compositions at Mars landing sites. Our results show that a significant fraction of weathering fluids loaded with Si, Fe, S, Mg, Ca, Cl, Na, K and Al remain in the liquid state at temperatures well below 273 K. We tested our model by analysing the mineralogies yielded by the evolution of the solutions: the resulting mineral assemblages are analogous to those actually identified on the Martian surface. This stability against freezing of Martian fluids can explain saline liquid water activity on the surface of Mars at mean global temperatures well below 273 K.

Was early Mars warm and wet, or cold and dry? Geomorphologies derived from liquid water flowing and ponding on the surface of the planet cover most of the Martian landscape[2], and there is much evidence of water-related mineralogies on the surface of Mars[3], implying that early Mars was wet. But several lines of evidence suggest that Mars has also been permanently cold[4], with mean global temperatures well below the freezing point of pure water. As a way of plausibly raising its surface temperatures above 273 K, climate models consider the addition of appreciable concentrations of CO_2 (ref. 11), CH_4 (ref. 5), NH_3 (ref. 1), CO_2-CH_4-NH_3-SO_2 (ref. 12) or SO_2-H_2S (ref. 13) to the early Mars atmosphere. Yet all these model scenarios have important limitations, as follows. Although geochemical models are compatible with low partial pressures of atmospheric CO_2 in early Mars[14], even a thick CO_2 atmosphere (5 bar) cannot efficiently raise the surface temperature above 273 K (ref. 6), and the greenhouse atmosphere produced by the CO_2-H_2O system would have saturated at temperatures well below 273 K (refs 5, 12). The amount of CH_4 required would imply a terrestrial-like biological source[15]. NH_3 would be consumed by ultraviolet photolysis in less than ten years[16]. Sulphur volatiles have brief residence times in the atmosphere[17], with no detectable sulphur-containing gases at present. Combinations of CO_2-CH_4-NH_3-SO_2 would provide only modest warrming[12]. Consequently, none of these models solves the problem of liquid water on a probably very cold early Mars. Here we address this early Mars climate paradox by instead modelling the stabilization against freezing of surface liquid solutions with large concentrations of dissolved solutes.

The reduction of the freezing point of aqueous solutions is a function of chemical composition and pressure. On Mars, large evaporitic deposits of sulphates[18] and chlorides[19], as well as hydrated and anhydrous salts and phyllosilicates[3], have been detected in numerous locations, showing that the composition of water-related minerals on the Martian surface must have derived from complex multicomponent solutions. More recently, geochemical models of early Mars aqueous solutions suggest salty rather than dilute fluids[20], and this salty nature of the Martian waters is a prerequisite for our model calculations (see Supplementary Information). To elucidate the thermodynamic behaviour of liquid solutions on early Mars, we have simulated the freezing and evaporation processes of fluids with a composition resulting from the weathering of basalts, as reflected in the chemical compositions at Mars landing sites[21] (Table 1). For all the modelled compositions, the initial solution salinities in the evaporation series (starting at 275 K)

Table 1 | Percentage element mass fraction at four Martian landing sites

Element	Viking 1 (Chryse Planitia)	Pathfinder (Ares Vallis)	Opportunity (Meridiani Planum)	Spirit (Gusev crater)
Al	4.4	4.5	4.69	4.6
Ca	4.6	4.8	4.53	4.07
Cl	0.71	0.49	0.47	0.72
Cr	-	0.16	0.3	0.23
Fe	13.2	14.6	13.8	11.8
K	0.22	0.63	0.34	0.28
Mg	3.3	4.2	4.18	5.02
Mn	-	0.4	0.27	0.24
Na	-	1.0	1.58	1.88
P	-	0.42	0.32	0.49
S	2.7	2.2	2.1	3.4
Si	22.6	21.0	20.2	18.4
Ti	0.41	0.55	0.68	0.59

[1]Space Science and Astrobiology Division, NASA Ames Research Center, Moffett Field, California 94035, USA. [2]Departamento de Geociencias Marinas, Universidad de Vigo, Lagoas Marcosende, Vigo 36200, Spain. [3]Centro de Astrobiología, CSIC-INTA, Torrejón de Ardoz 28850, Madrid, Spain. [4]Centro de Biología Molecular Severo Ochoa, CSIC-UAM, Cantoblanco 28049, Madrid, Spain.

range between 5 and 6% (based on samples taken by the following Mars landers: 51.8 g l^{-1} Pathfinder, 54.21 g l^{-1} Viking, 54.21 g l^{-1} Spirit, 57.93 g l^{-1} Opportunity), only slightly higher than terrestrial sea water (~3.51%). This is because we impose an equilibrium between Si and amorphous SiO_2, and so the excess of Si immediately precipitates as amorphous silica until under-saturation is reached. The remaining Si is ultimately involved in the precipitation of phyllosilicates. As the model system evolves, the dropping temperature and the freezing concentration increase the salinities in all cases (Table 2). Even at the high salinities reached at very low temperatures, the solutions still have a perfectly liquid appearance and behaviour, consistent with the shape and morphology of the valley networks, outflow channels and gullies seen on Mars, which suggest rapid fluvial flow rather than the slow, halting movement of thick eutectic brines with a high density and viscosity.

A certain degree of atmospheric pressure was necessary both to achieve surface temperatures that allow surface solutions to flow (otherwise global mean temperatures would have always been below 200 K, similar to those today) and to lower the vapour pressure of the liquid solutions with respect to the atmospheric pressure (so that they can exist as liquids without evaporating and becoming crusts of hydrated materials). We assumed a CO_2 atmosphere with an average pressure on the surface of 2 bar. This atmospheric partial pressure p_{CO_2} gives surface temperatures near 245 K (ref. 6), which is enough warming to enable the flow of the solutions modelled here. However, our model results are independent of the initial p_{CO_2} in the atmosphere. The same model run with an atmosphere of 10 mbar of CO_2 (which is above the triple point of water, 6.11 mbar) results in very similar outcomes in terms of liquid water stability, and shows only slight differences with respect to the sequence of mineral phase precipitation.

Our goal was to learn how a combination of different processes of evaporation and freezing affect the lowering of the freezing-point temperature of a hypothetical Martian solution. Additionally, our model calculations helped to clarify the sequence of phase formation and destabilization during the temporal evolution of the solution. We developed equilibrium models obtained by inducing lineal processes of evaporation and freezing of the liquid solutions (see Methods section). At subzero temperatures, evaporation and freezing are very efficient at increasing the ionic activity and lowering the freezing point of the solution. Our results on the evolution of residual liquid water and the phase transitions as a function of temperature are presented in Fig. 1. In all modelled scenarios, a significant fraction of liquid water remains stable at temperatures below 273 K. For example, at temperatures of 263, 250, 245 and 223 K, up to 78%, 22%, 14% and 6% of the original water reservoir remains in the liquid state, respectively. A direct comparison can be done with what is expected to occur in terrestrial sea water, where at 263 and 223 K, 20% and 0.31% of the original water would remain unfrozen[22]. As a result of liquid water loss through freezing and evaporation, a sequence of phase transitions dependent on the saturation equilibrium of the different components in the solution is observed. The sequence of phase precipitation and the freezing point depression both follow the same general trends for each of the landing sites' ionic composition, although minor variations can be observed by introducing changes in the initial ionic concentrations. We also studied how these ionic changes can affect the concrete behaviour of the solutions, and our results are detailed in the Supplementary Information.

Table 2 | Effect of freezing concentration (Pathfinder data)

Temperature (K)	Mass of water (kg)	Salinity (g l^{-1})	Water activity
275	1.0	51.8	0.99
263	2.171×10^{-1}	83.91	0.98
253	5.721×10^{-2}	317.0	0.97
233	2.350×10^{-2}	879.9	0.87

In general, a very fast evaporation process would induce a lowering of the freezing point. This is because when the evaporation rate increases with respect to freezing, a smaller fraction of liquid water remains at 250 K, as some water is lost in the evaporation process and in the precipitation of hydrated phases, further concentrating the dissolved species. A fast evaporation yielded the largest drop in freezing point, down to 220 K. But, at the same time, the saturation of

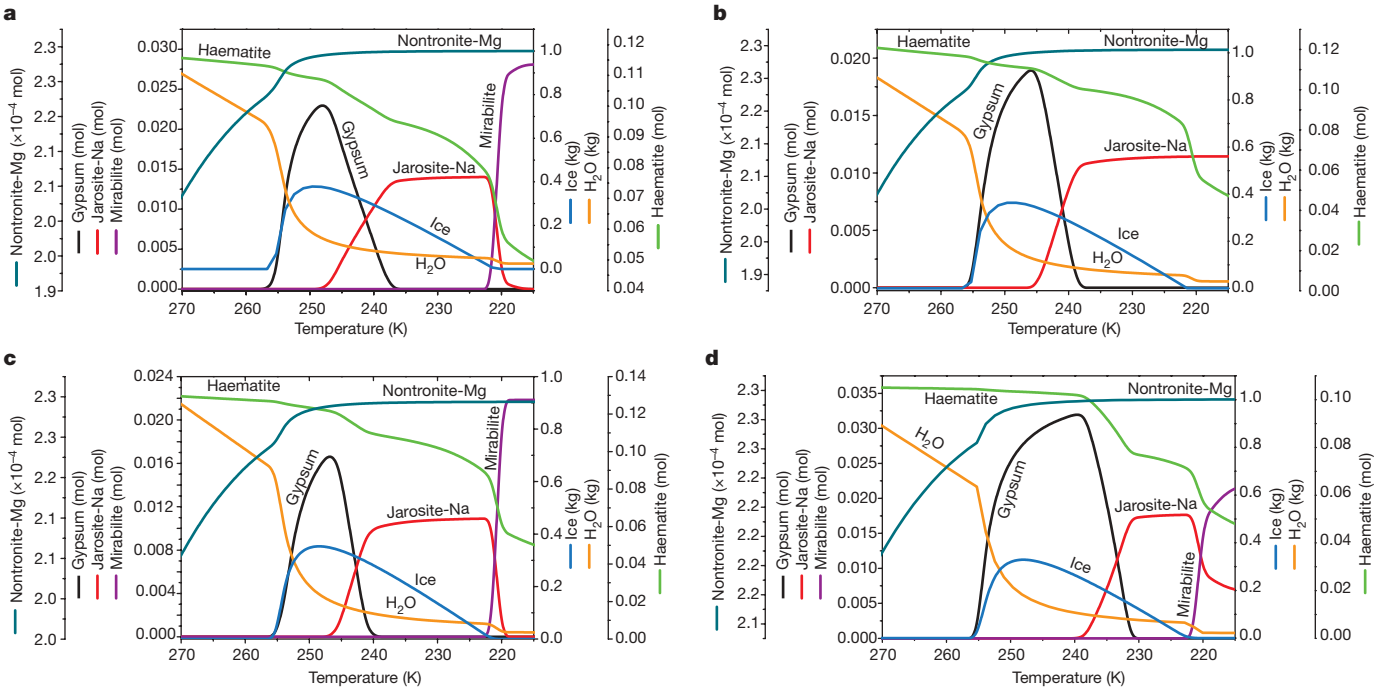


Figure 1 | Liquid water on Mars at subzero temperatures. Residual water mass, ice formation and change in ion concentrations as a function of temperature in four hypothetical Martian solutions based on data from four Mars landers (Table 1), in a model of evaporation and freezing, followed by continuing freezing down to 215 K. Note the opposite behaviour for the different evaporitic phases (mainly gypsum versus jarosite) and their influence on the modification of the available liquid water mass. **a**, Viking; **b**, Pathfinder; **c**, Opportunity; **d**, Spirit.

many salts increases rapidly, promoting phase precipitation and therefore lowering the concentration of dissolved ions, which favours the formation of ice.

To understand evaporite composition and liquid lines of descent for Mars, we analysed the rate of ice formation in different situations, varying the relative velocity of evaporation versus freezing, and the total number of moles of water evaporated (Fig. 2). For early Mars, the preferred scenarios are those where evaporation occurred at moderate velocities (20 or 30 steps in Fig. 2a), and in which a compromise between evaporation, precipitation and freezing rates is reached, and therefore wherein the maximum amount of free liquid water at the minimum possible temperatures is attained. The minimum quantity of liquid water shown in Fig. 2 could be as voluminous as required, depending on an increase in the initial water mass in the system, and an increase in the initial ionic concentration, keeping the relative proportions in weight, which would result in brine formation at the beginning of the process. Therefore, the higher the initial water mass on Mars and its ionic concentration, the higher the mass of the water that would remain in the liquid state at very low ambient temperatures. Also, higher initial water masses will form more dilute solutions, and less initial water will create concentrated brines. Our results predict the precipitation of phyllosilicates starting at 265 K, and therefore at high water activities; while the formation of sulphates would start around 245 K, a situation in which ice would be the dominant water phase in the system. This outcome may reflect a scenario in which initial solutions favourable to the formation of phyllosilicate-type minerals in the Noachian period evolved as Mars gradually dried out into hypersaline solutions more conducive to the synthesis of sulphate-type deposits during the Late Noachian to Hesperian periods[3]. Alternatively, all these mineralogies could be a result of the contemporary evolution of different aqueous solutions under locally diverse environmental conditions.

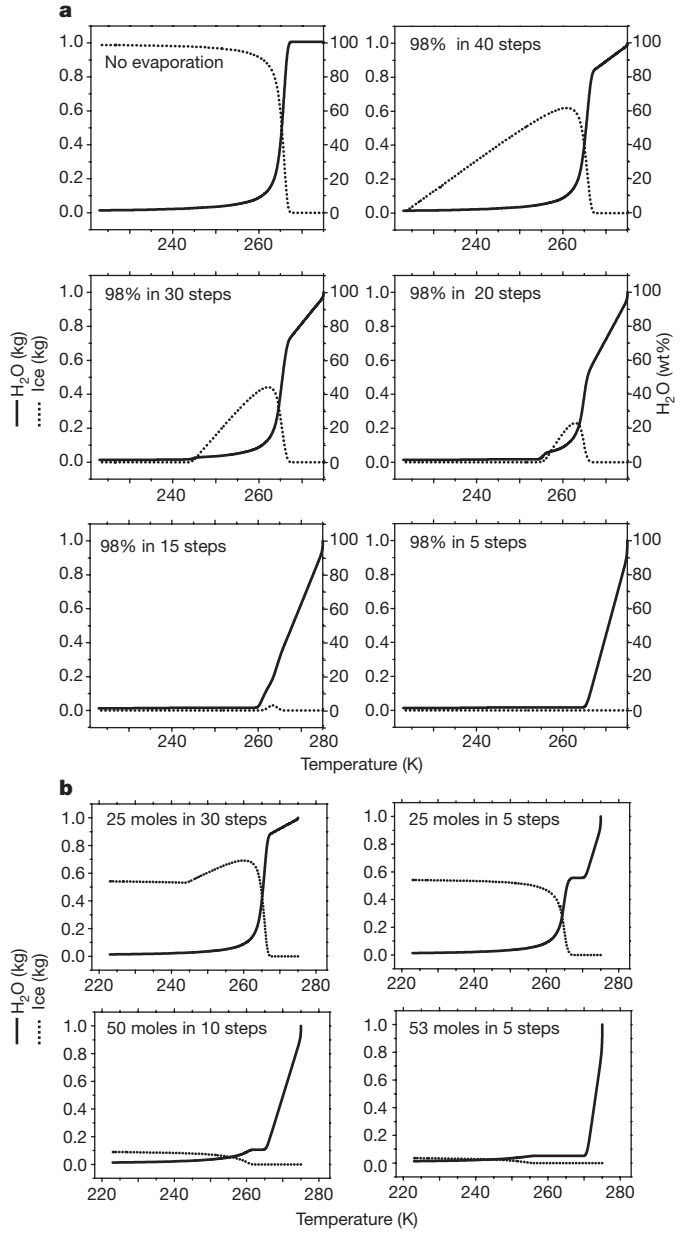


Figure 2 | Analysis of ice formation during the evaporation/freezing sequence. Evaporation of 98% of the initial liquid water mass at different velocities, related to the same freezing process, assuming that supersaturated salts were not allowed to precipitate and remained metastable. Ion composition is that in Table 1 for the Opportunity landing site. **a**, Water mass is 1 kg. The scenario of no evaporation hints at the maximum possible freezing before ice starts to form. In the other scenarios the evaporation velocity is progressively faster with respect to freezing. As the water undergoes evaporation up to 98%, the final masses of ice and water are very small. In the bottom panels, ice formation is barely observed, and the water curve (although the water mass is very small) is always over the ice curve. **b**, Different scenarios in which we varied both the relative velocity of evaporation versus freezing and the total number of moles of water evaporated.

Although saline solutions may not be responsible for the formation of all of the fluvial features on Mars, the geochemical analysis we present does help to explain the long-term presence of open bodies of salty water on a cold Mars. A combination of atmospheric warming by a mixture of greenhouse gases due to extensive volcanism[23] and obliquity controls on climate[24], resulting in surface temperatures around 245 K (ref. 6), and melting point depression of the aqueous solutions via ionic loading that stabilized the liquids at temperatures below 273 K, would have led to the flow of liquid water on the Martian surface. This model combination was only achieved during the early history of Mars, before the significant atmospheric loss that occurred by the end of the Noachian by hydrodynamic and impact escape[25].

However, water-shaped features have formed throughout Martian history, and may even possibly do so today to some extent[26]. Our results are as relevant to recent processes as to early Mars, because qualitatively similar results hold when a few tens of millibars or less of CO_2 are considered. During the Amazonian period, it is more about maintaining stability against desiccation and salt crust formation, keeping the non-eutectic-acid solutions stable as liquids on the surface. Alternatively, the aqueous solutions could be extant beneath the surface at shallow depths and erupt through permafrost. In the Noachian and the Hesperian periods, less than half of the initial liquid water reservoir of Mars would have been lost through evaporation or by entrapment in the crystal structure of the precipitated hydrated salts, further supporting the idea that most of the water on the Martian surface consisted of supercooled water solutions shaping valleys and seas, partly covered by large masses of ice. In such aqueous solutions, the activity of water (a thermodynamic property of solutions that is dependent on temperature and salt concentration) would have almost never reached values below 0.95 (Table 2), and would thus be very unlikely to challenge any possible early Martian biosphere, as is proposed to have occurred at temperatures above freezing[20]. Our results are compatible with Mars lander and orbiter data and with climate modelling, and suggest a cold and wet early Mars.

METHODS SUMMARY

The geochemical calculations were performed using the Phreeqc2.15.0 software[27]. The model was run at decreasing temperatures from 270 to 215 K and with the following initial conditions: H_2O = 54 mol (1 kg), p_{CO_2} = 2 bar, pH = 4.0, redox potential Eh = 4.0, all sulphur as S^{6+}, all iron as Fe^{3+}. As shown in the Supplementary Information, these initial values are not determinant for the general behaviour of liquid water and ice. The evolution of the system was

analysed at specific temperature steps (40 steps, $\Delta T = 1.375$ K, see Supplementary Information), allowing solution/solid-phase equilibration at each one of these steps. For the overall evaporation and cooling calculations, the temperature dependence of the equilibrium constant was modelled using the Van't Hoff expression. The EQ3/6 database provided by the Lawrence Livermore National Laboratory[27] was employed for the calculations. In the version we used, the Phreeqc program implements both the ion-pairing based on the Debye–Hückel approach for the calculation of activities as the product of several fundamental constants[28], and the Pitzer virial-coefficient model for highly saline solutions or brines[29]. The general system behaviour was modelled with the Debye–Hückel database to obtain the evaporation and freezing guidelines, including silicate behaviour. The final stages of the evaporation process were modelled using the Pitzer database because it is the better approach for high ionic strengths, although it has two main drawbacks: the lack of interaction parameters for aqueous Al and Si species, which precludes any calculations with aluminosilicates; and the very few data available for redox reactions. In our case, Fe^{3+} values were added to the original Pitzer database by using recent compilations[20,30].

Received 19 August 2008; accepted 16 March 2009.

1. Sagan, C. & Mullen, G. Earth and Mars: evolution of atmospheres and surface temperatures. *Science* **177,** 52–56 (1972).
2. Baker, V. R. Water and the martian landscape. *Nature* **412,** 228–236 (2001).
3. Bibring, J. P. *et al.* Global mineralogical and aqueous Mars history derived from OMEGA/Mars Express data. *Science* **312,** 400–404 (2006).
4. Gaidos, E. & Marion, G. Geological and geochemical legacy of a cold early Mars. *J. Geophys. Res.* **108,** doi:10.1029/2002JE002000 (2003).
5. Kasting, J. F. CO_2 condensation and the climate of early Mars. *Icarus* **94,** 1–13 (1991).
6. Colaprete, A. & Toon, O. B. Carbon dioxide clouds in an early dense Martian atmosphere. *J. Geophys. Res.* **108,** 5025, doi:10.1029/2002JE001967 (2003).
7. Griffith, L. L. & Shock, E. L. Hydrothermal hydration of Martian crust: illustration via geochemical model calculations. *J. Geophys. Res.* **102,** 9135–9143 (1997).
8. Segura, T., Toon, O. B., Colaprete, A. & Zahnle, K. Environmental effects of large impacts on Mars. *Science* **298,** 1977–1980 (2002).
9. Brass, G. W. Stability of brines on Mars. *Icarus* **42,** 20–28 (1980).
10. Kuzmin, R. O. & Zabalueva, E. V. On salt solutions in the Martian criolithosphere. *Solar Syst. Res.* **32,** 187–197 (1998).
11. Forget, F. & Pierrehumbert, R. T. Warming early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278,** 1273–1276 (1997).
12. Squyres, S. W. & Kasting, J. F. Early Mars: how warm and how wet? *Science* **265,** 744–749 (1994).
13. Halevy, I., Zuber, M. T. & Schrag, D. P. A sulfur dioxide climate feedback on early Mars. *Science* **318,** 1903–1907 (2007).
14. Chevrier, V., Poulet, F. & Bibring, J.-P. Early geochemical environment of Mars as determined from thermodynamics of phyllosilicates. *Nature* **448,** 60–63 (2007).
15. Kasting, J. F. Warming early Earth and Mars. *Science* **276,** 1213–1215 (1997).
16. Kuhn, W. R. & Atreya, S. W. Ammonia photolysis and the greenhouse effect in the primordial atmosphere of the Earth. *Icarus* **37,** 207–213 (1979).
17. Johnson, S. S., Mischna, M. A., Grove, T. L. & Zuber, M. T. Sulfur-induced greenhouse warming on early Mars. *J. Geophys. Res.* **113,** E08005, doi:10.1029/2007JE002962 (2008).
18. Gendrin, A. *et al.* Sulfates in martian layered terrains: the OMEGA/Mars Express view. *Science* **307,** 1587–1591 (2005).
19. Osterloo, M. M. *et al.* Chloride-bearing materials in the southern highlands of Mars. *Science* **319,** 1651–1654 (2008).
20. Tosca, N. J., Knoll, A. H. & McLennan, S. M. Water activity and the challenge for life on early Mars. *Science* **320,** 1204–1207 (2008).
21. Karunatillake, S. *et al.* Chemical compositions at Mars landing sites subject to Mars Odyssey Gamma Ray Spectrometer constraints. *J. Geophys. Res.* **112,** E08S90, doi:10.1029/2006JE002859 (2007).
22. Marion, G. M. & Kargel, J. S. *Cold Aqueous Planetary Geochemistry with FREZCHEM* 102–109 (Springer, 2008).
23. Phillips, R. J. *et al.* Ancient geodynamics and global-scale hydrology on Mars. *Science* **291,** 2587–2591 (2001).
24. Sagan, C., Toon, O. B. & Gierasch, P. J. Climatic change on Mars. *Science* **181,** 1045–1049 (1973).
25. Brain, D. A. & Jakosky, B. M. Atmospheric loss since the onset of the Martian geologic record: combined role of impact erosion and sputtering. *J. Geophys. Res.* **103,** 22689–22694 (1998).
26. Fairén, A. G. *et al.* Evidence for Amazonian acidic liquid water on Mars—A reinterpretation of MER mission results. *Planet. Space Sci.* **57,** 276–287 (2009).
27. Parkhurst, D. L. & Appelo, C. A. J. *User's Guide to PHREEQC (Version 2)—A Computer Program for Speciation, Batch-Reaction, One-Dimensional Transport, and Inverse Geochemical Calculations* US Geol. Surv. Wat. Resour. Invest. Rep. 99–4259 (1999).
28. Debye, P. & Hückel, E. Zur theorie der electrolyte: I. Gefrierpunkterniedindung und Verwandte Ersheinungen. *Phys. Z.* **24,** 185–206 (1923).
29. Pitzer, K. S. *Thermodynamics* 3rd edn (McGraw-Hill, 1995).
30. Marion, G. M., Kargel, J. S. & Catling, D. C. Modeling ferrous-ferric iron chemistry with application to Martian surface geochemistry. *Geochim. Cosmochim. Acta* **72,** 242–266 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Work by A.G.F. and A.F.D. was supported by ORAU-NPP. We thank J. Kasting and J. Kargel for reviews that significantly improved the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.G.F. (alberto.g.fairen@nasa.gov).

LETTERS

Breakdown of the Bardeen–Cooper–Schrieffer ground state at a quantum phase transition

R. Jaramillo[1], Yejun Feng[1,2], J. C. Lang[2], Z. Islam[2], G. Srajer[2], P. B. Littlewood[3], D. B. McWhan[4] & T. F. Rosenbaum[1]

Advances in solid-state and atomic physics are exposing the hidden relationships between conventional and exotic states of quantum matter. Prominent examples include the discovery of exotic superconductivity proximate to conventional spin and charge order[1,2], and the crossover from long-range phase order to preformed pairs achieved in gases of cold fermions[3–5] and inferred for copper oxide superconductors[5]. The unifying theme is that incompatible ground states can be connected by quantum phase transitions. Quantum fluctuations about the transition are manifestations of the competition between qualitatively distinct organizing principles[6,7], such as a long-wavelength density wave and a short-coherence-length condensate. They may even give rise to 'protected' phases, like fluctuation-mediated superconductivity that survives only in the vicinity of an antiferromagnetic quantum critical point[8,9]. However, few model systems that demonstrate continuous quantum phase transitions have been identified, and the complex nature of many systems of interest hinders efforts to more fully understand correlations and fluctuations near a zero-temperature instability. Here we report the suppression of magnetism by hydrostatic pressure in elemental chromium, a simple cubic metal that demonstrates a subtle form of itinerant antiferromagnetism[10–16] formally equivalent to the Bardeen–Cooper–Schrieffer (BCS) state in conventional superconductors. By directly measuring the associated charge order in a diamond anvil cell at low temperatures, we find a phase transition at pressures of ~10 GPa driven by fluctuations that destroy the BCS-like state but preserve the strong magnetic interaction between itinerant electrons and holes. Chromium is unique among stoichiometric magnetic metals studied so far in that the quantum phase transition is continuous, allowing experimental access to the quantum singularity and a direct probe of the competition between conventional and exotic order in a theoretically tractable material.

Chromium is a b.c.c. crystal that forms a spin-density wave (SDW) below a Néel temperature of $T_N = 311$ K. This weak-coupling ground state of electron–hole pairs is described by a BCS-like exponential expression for the magnetic moment, μ, of the form $\mu(T=0) \propto \exp(-1/\lambda)$, where the coupling constant, λ, can be tuned by application of pressure, P, and chemical doping. The SDW is modulated by a wavevector, **Q**, of magnitude Q (in units of $2\pi/a$, where a is the lattice constant), which is selected by the nesting condition and is slightly incommensurate with the crystal lattice (Fig. 1, inset). The SDW is accompanied by an itinerant charge density wave (CDW), which is modulated by 2**Q** and is thought of as the second harmonic of the SDW[17]. This harmonic relationship between spin and charge is consistent with the $I_{CDW} \propto I_{SDW}^2 \propto \mu^4$ scaling (where I is scattering intensity), observed as a function of both temperature[18] and pressure[14].

Tuning the magnetism towards the quantum phase transition using pressure while directly measuring the charge and spin order parameters is the conceptually preferred route to studying the quantum critical regime. By using P as our tuning variable, we avoid the complications of chemical doping, such as variable band filling and substitutional disorder, that may be significant at the critical point. Our high-resolution X-ray diffraction technique provides a direct measurement of the order parameters and a view into the microscopic world of nested electron–hole pairs. However, the

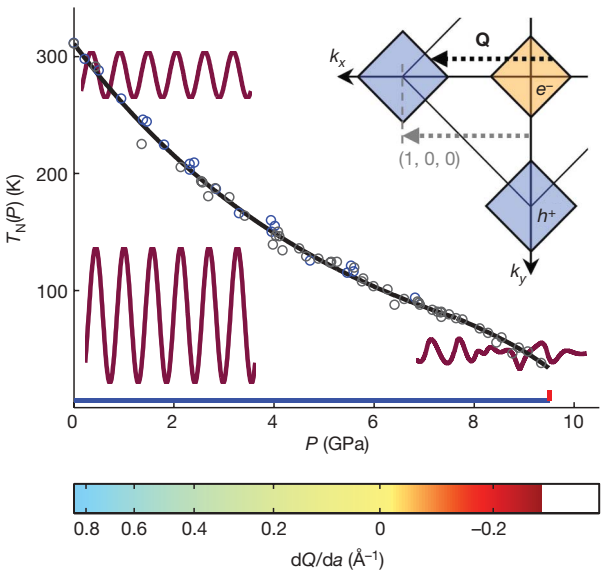


Figure 1 | Phase diagram and Brillouin zone schematic for magnetism in chromium. Phase diagram shows $T_N(P)$; grey points are determined from our data using the mean-field relation $T_N \propto I_{CDW}^{1/4}(T=0)$, blue points are determined from resistivity measurements of T_N (ref. 11; also see refs 13, 14 and Fig. 4 for a discussion of pressure scales) and black line is a guide to the eye. Blue and red lines indicate experimental cuts (see Supplementary Information for data collected along the red trajectory). The colour bar charts the change in SDW wavevector, Q, with pressure (at $T < 8$ K) as measured by dQ/da, the derivative with respect to chromium lattice constant. The quantum critical regime is marked by both a deviation from the exponentially tuned BCS ground state and an unexpected changed in the sign of dQ/da. The three cartoons are proposed representations of the density-wave order parameter in three regimes. At low T and low P, the order-parameter amplitude is large and highly ordered. At high T and low P, the amplitude is reduced before being cut off by a first-order phase transition. At low T and high P, the amplitude is further reduced and the transition is driven by transverse fluctuations (not drawn to scale). Inset, schematic of the first Brillouin zone, showing the incommensurate wavevector, **Q**, connecting the nested sections of Fermi surface that are eliminated in the magnetic phase by the formation of an exchange-split energy gap[16].

[1]The James Franck Institute and Department of Physics, The University of Chicago, Chicago, Illinois 60637, USA. [2]The Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA. [3]Cavendish Laboratory, University of Cambridge, Cambridge CB3 OHE, UK. [4]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

experimental challenges are daunting. The pressure required to drive the transition is of the order 100,000 atm, necessitating the use of diamond-anvil-cell and synchrotron X-ray diffraction techniques. Tracking just the CDW order parameter into the quantum critical regime requires measuring satellite diffraction peaks that are nearly 10^{10} times weaker than the lattice Bragg reflections in high-quality single-crystal samples at high pressure and low temperature.

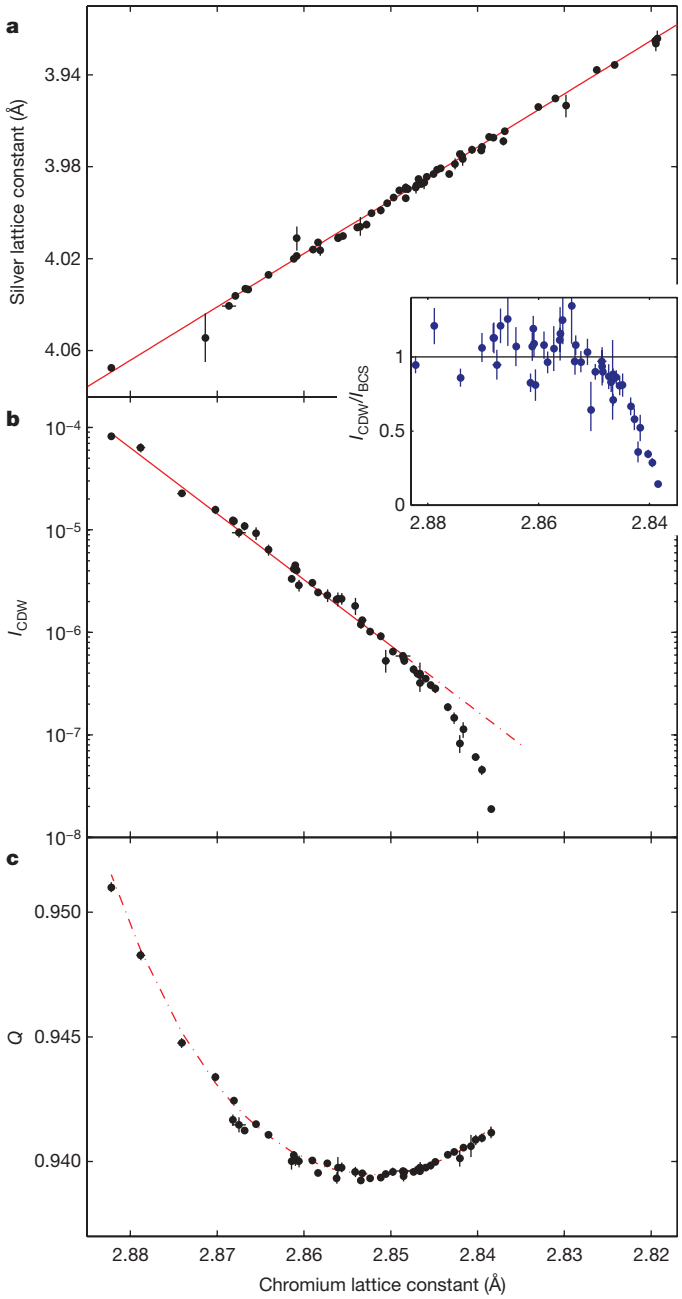


Figure 2 | Structure, CDW intensity and wavevector *Q* for $T < 8$ K. **a**, Variation of the silver lattice constant over a wide range of chromium lattice constant. The high-fidelity linear fit rules out a significant first-order transition in the critical pressure range. **b**, Normalized CDW intensity, I_{CDW}, summed over all three domain types (see Methods); for $a = 2.8348$ Å, we have established an upper bound on I_{CDW} of 5.9×10^{-9} (not plotted). The exponential fit to the data in the non-critical regime is shown in red. Inset, fractional deviation of I_{CDW} from the mean-field ground state, I_{BCS}, determined from the exponential fit in **b**. **c**, Magnitude, *Q*, of the SDW wavevector, showing an unexpected upturn in the critical pressure range. This contrasts with the gently decreasing dependence on pressure in the non-critical regime. The red curve is a polynomial fit used to generate d*Q*/d*a* as presented in Fig. 1. Error bars on lattice constants and *Q* represent s.d.; error in I_{CDW} is s.e.m.

In Fig. 2, we present X-ray diffraction measurements of the CDW order parameter, the magnitude of magnetic wavevector, *Q*, and the chromium lattice in the quantum critical regime. Our data establish the existence of a continuous, antiferromagnetic quantum phase transition in this stoichiometric model system. In Fig. 2a, we plot the lattice constant of silver, our pressure calibration, over a wide range of chromium lattice constant corresponding to $0 < P < 14.5$ GPa. The continuous evolution of the chromium lattice throughout this pressure range precludes the existence of a significant first-order quantum phase transition, as has been seen in direct order-parameter measurements for every other stoichiometric, itinerant magnet studied so far[19].

The evolution of the CDW order parameter through this pressure range is presented in Fig. 2b, where we plot the CDW diffraction intensity, I_{CDW} (see Methods), as a function of chromium lattice constant. I_{CDW} decreases by more than two orders of magnitude between 0 and 7 GPa, where it is faithfully described by an exponential dependence on *P* (refs 14, 15). Above 7 GPa, the intensity falls away from this mean-field expectation, and by 9.5 GPa it is nearly one order of magnitude weaker than the extrapolated exponential fit (Fig. 2, inset). At the highest measured pressure, the magnetic moment satisfies $\mu/\mu_0 = (I_{CDW}/I_{CDW,0})^{1/4} = 0.12$, where $\mu_0 = 0.41\mu_B$ is the root-mean-square ordered moment at ambient pressure and low temperature, μ_B is the Bohr magneton and $I_{CDW,0}$ is the CDW diffraction intensity at zero pressure and base temperature. This should be compared with the value $\mu/\mu_0 = 0.27$, which obtains at the weakly first-order Néel transition at ambient pressure[20]. The antiferromagnetic phase transition therefore crosses from first-order to second-order behaviour as we pass from the thermal to the quantum regime.

The deviation of I_{CDW} from the weak-coupling ground state is accompanied by an unexpected upturn in *Q* (Fig. 2c and the colour bar in Fig. 1). In the exponentially tuned regime, *Q* decreases slowly with pressure (by sharp contrast with the response to chemical doping[15]) and appears to level off for $P > 4$ GPa, providing evidence for a rigid band structure in chromium at this pressure scale[14,15]. The upturn in *Q* above 7 GPa signals the start of a new relationship between pressure, magnetism and band structure.

To understand the origin of this behaviour, we first turn to a mean-field theory that is appropriate for the system at ambient pressure. In the absence of fluctuations, the free energy of the magnetic state can be written as

$$F = \frac{1}{2}a_0|\psi_0|^2 + \frac{1}{4}b_0|\psi_0|^4 + \frac{1}{2}\xi_0^2(\nabla\psi_0)^2 + \frac{1}{2}e_0|\psi_0|^2|\nabla\phi - \mathbf{q}|^2 + f_0|\psi_0|^4\cos(2\phi) \quad (1)$$

where $|\psi_0|e^{i\phi}$ is the SDW order parameter, $\mathbf{q}$ is the nesting vector as determined by the band structure and the coefficients a_0, b_0, …, f_0 characterize the SDW state. The final two terms allow $Q = |\nabla\phi|$ to vary from $q = |\mathbf{q}|$, the first expressing the cost of repopulating the Brillouin zone to accommodate $Q \neq q$, and the second reflecting the coupling of the CDW to the lattice.

The different ψ_0 dependences of the final two terms equation (1) allow us to distinguish between the various mechanisms that affect *Q*. In Fig. 3a, we show *Q* as a function of the ordered moment $\mu \cong |\psi_0|$ for both the temperature- and the pressure-driven phase transitions. In the thermal case, *Q* varies rapidly when μ is strong ($\mu \approx \mu_0$), but levels off for $\mu/\mu_0 < 0.6$. This identifies $0.6 < \mu/\mu_0 < 1$ as the regime in which the final two terms in equation (1) are in competition; for smaller magnetic moments, the quartic term is too weak to affect *Q*. For the pressure-driven transition, the regime $0.6 < \mu/\mu_0 < 1$ is likewise characterized by a monotonic change of *Q* with μ, with *Q* levelling off as the magnetism is further suppressed, before suddenly increasing for $\mu/\mu_0 < 0.3$. This upturn on the approach to the critical regime cannot be explained by competition between the two *Q*-dependent terms in equation (1). We measured the temperature dependence of *Q* at $P = 9.5$ GPa and found it to be constant up to $T = 15$ K (see experimental cut in Fig. 1 and data in Supplementary Information), ruling out the effects of finite

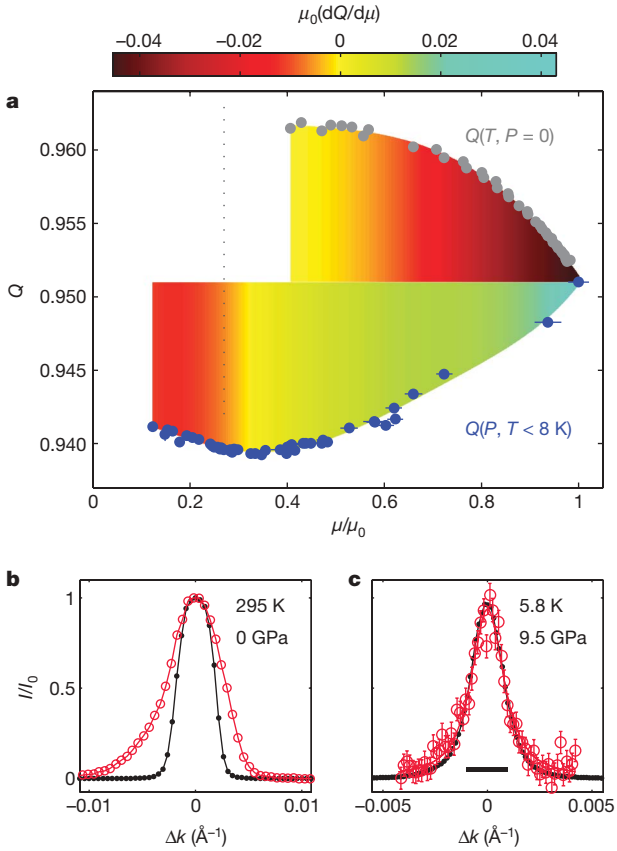


Figure 3 | Dependence of Q on order parameter and band structure in the classical and quantum regimes. **a**, Variation of Q with ordered moment, μ, for both temperature- and pressure-driven phase transitions. Colour maps represent $\mu_0(\mathrm{d}Q/\mathrm{d}\mu)$ determined by smooth polynomial fits to the data in the thermal (upper colour map) and quantum (lower colour map) cases. Data for $Q(T, P = 0)$ are taken from the literature[18,29] and $\mu(T)$ is established by neutron diffraction[20]. For our high-pressure measurements, μ is calculated using the $\mu \propto I_{\mathrm{CDW}}^{1/4}$ mean-field relationship. The vertical dotted line marks the level at which the magnetic order is cut off by a weak first-order phase transition at the ambient pressure Néel temperature. Error bars on Q represent s.d.; error in μ is s.e.m. **b**, **c**, Longitudinal CDW (red) and lattice (black) diffraction line shapes in the thermal (**b**) and quantum (**c**) critical regimes. All scans are normalized to peak intensity, and error bars are s.e.m. In the thermal case, the lattice and CDW diffraction peaks are $(2, 0, 0)$ and $(4 - 2Q, 0, 0)$, and in the quantum case they are $(2, 1, 1)$ and $(4 - 2Q, 1, 1)$; these reflections are chosen to optimize the structure factor in the critical regime[15]. Asymmetric broadening of the CDW line shape near the thermal transition reflects imperfect Fermi surface nesting at high temperature[15]. By contrast, the CDW line shape near the quantum transition is instrument-limited, indicating that the Fermi surface nesting remains excellent. Instrument resolution (FWHM) is indicated by the black bar; the instrument-limited line shape places a lower bound of 2,000 Å on the longitudinal CDW correlation length.

temperature as an explanation. Although we cannot rule out changes to the underlying band structure, a non-monotonic evolution of q with a reduction in a of only 1.5% seems unlikely, as is a coincidence of the upturn in Q with the approach to the critical regime. We conclude that equation (1) is inadequate for describing the evolution of Q in the quantum critical regime, and instead require an expression that goes beyond the mean-field assumption to include a full set of coupled fluctuation terms at second order in ψ_0.

Although Q responds to changes in q, it does not reflect more subtle changes to the band structure that can affect the magnetism. The nesting feature of the paramagnetic band structure leads to a peak in the wavevector (k)-dependent non-interacting susceptibility, $\chi_0(k)$, at $k = q$. For ideal nesting (perfectly flat Fermi surfaces) this peak diverges logarithmically and the weak-coupling ground state is stable for arbitrarily small coupling constant, λ. As a consequence, a_0, the

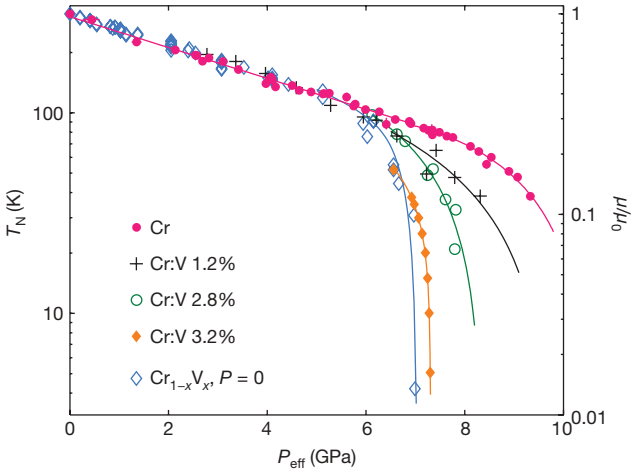


Figure 4 | Disparate routes to quantum criticality. The effects of pressure and vanadium doping on the magnetic ordered moment, μ (or T_N), are plotted together to reveal a family of antiferromagnetic quantum phase transitions. By compressing pure chromium into the critical regime, we are able to distinguish the effects of pressure and chemical doping on the magnetism (compare with fig. 4 in ref. 13). Data for $Cr_{1-x}V_x(P = 0)$ and results for $x = 1.2\%$, 2.8% and 3.2% under pressure are taken from the literature[13–15,30]. All curves are determined from measurements of T_N except for our scattering data, for which the measured quantity is $I_{\mathrm{CDW}} \propto \mu^4$. Pressure and doping are related by 2.05 GPa = 1% V, set by the collapse of the data for $I_{\mathrm{CDW}}^{1/4}(P < 7\ \mathrm{GPa}, x = 0)$ and $T_N(P = 0, x < 2.5\%)$ onto the same exponential curve. Data are plotted against effective pressure, $P_{\mathrm{eff}} = P + (2.05\ \mathrm{GPa})100x$, the sum of applied and chemical pressure. The reported pressure scale for the data for vanadium doping of 1.2% and 2.8% has been reduced by a factor of 1.2. This reduction factor collapses the reported $T_N(P, x = 0)$ curve onto our measured $I_{\mathrm{CDW}}^{1/4}(P, x = 0)$ curve, where the former was measured by the same group using the same experimental technique as here for $T_N(P, x = 1.2\%)$ and $T_N(P, x = 2.8\%)$ (refs 11, 14, 30). Disorder destabilizes the BCS ground state but, unlike for pure chromium, does not permit deconvolution of the physics of fluctuations, impurity scattering and Fermi surface warping.

quadratic coefficient in equation (1), vanishes non-analytically at a critical mean-field pressure P_{C0} (in fact $a_0^{\mathrm{BCS}} \propto \exp[1/(P - P_{C0})]$) and the correlation length, $\xi = \xi_0(a_0)^{-1/2}$, diverges so rapidly that quantum fluctuations are irrelevant in the BCS theory. Under realistic nesting conditions, the peak at $\chi_0(k = q)$ is finite and the weak-coupling theory is more limited in scope. The critical point now occurs at a non-zero coupling constant, λ_C, and the conventional Ginzburg–Landau behaviour is recovered: $a_0^{\mathrm{MF}} \propto P - P_{C1}$. Hence, a renormalized transition (but one still mean-field-like in this approximation) at a lower critical pressure, $P_{C1} < P_{C0}$, is expected. Crucially, the correlation length vanishes as a power law and quantum fluctuations will play a role close to the true critical point.

The nesting condition is characterized by the warping of the magnetic Fermi surfaces away from being ideally flat. This warping is measured by the longitudinal CDW line scans presented in Fig. 3. Close to the ambient-pressure Néel transition, the CDW line shapes[15] are broad and asymmetric (Fig. 3b). The broadening is a straightforward result of thermal melting of the density wave[18], and the asymmetry results from warped Fermi surfaces, which lead to an asymmetric set of low-energy excitations away from the equilibrium Q value. The longitudinal CDW line scans remain broadened even at 5 K at ambient pressure, where the (deconvolved) CDW full-width at half-maximum (FWHM) is 1×10^{-3} Å^{-1}, pointing to the presence of longitudinal CDW excitations ($\Delta Q/Q = 0.02\%$) even at this low temperature. By contrast, a longitudinal CDW line scan made at 9.5 GPa and 6 K (Fig. 3c) shows no signs of broadening to within 3×10^{-4} Å^{-1}. Q is stiffer in the longitudinal direction at $T = 6$ K and $P = 9.5$ GPa than at 5 K and 0 GPa, indicating an improvement in the nesting condition under pressure. We therefore consider the quantum phase transition not to

be a failure of nesting, but rather to be a crossover into a regime in which weak-coupling theory is invalid despite the existence of a well-defined nesting feature[21].

To better understand this crossover, in Fig. 4 we consider the family of antiferromagnetic quantum phase transitions that is related by varying applied pressure and vanadium doping. It is immediately apparent that pressure and doping cannot be brought into a one-to-one correspondence throughout the phase diagram. This suggests different microscopic roles for these two tuning variables, consistent with the qualitatively distinct dependences Q has on pressure and doping[15]. The effect of chemical doping on the nested band structure is underscored by recent results[22] that suggest multiple $\mathbf{Q}$ vectors in highly-doped Cr:V, where no such evidence exists for the pure system.

We can discuss these differences within a single conceptual framework by considering the relevant length scales on which the ground state deviates from the weak-coupling prediction. The electron–hole-pair coherence length for the parent system (pure chromium at zero temperature and pressure) is $\xi \approx v_F/\Delta = 35$ Å, where v_F is the Fermi velocity. In general, weak-coupling theory should become more appropriate as the gap, Δ, is suppressed: to escape the weak-coupling ground state, we require a cut-off length scale beyond which pair coherence breaks down. For doping-driven transitions, both impurity scattering and changes to the nesting condition can provide this length scale, and both agree qualitatively with the systematic trend in the critical pressure with alloying. The observation of non-mean-field scaling in the excess magnetic resistivity at the pressure-driven quantum phase transition in Cr:V 3.2% points to the importance of fluctuations in the quantum regime[13]. The difficulty lies in relating this transport signature of quantum criticality to the requisite cut-off length scale, and in distinguishing between multiple sources for this cut-off. Importantly, and by contrast, this complication is absent for pure chromium under pressure, where fluctuations provide the most natural explanation for a breakdown length scale. This highlights the advantages of studying the parent system with a clean tuning parameter, including the need for magnetotransport and thermodynamic measurements at the dissolution of the BCS state, and provides access to a regime of longer pair coherence lengths, where lower-energy fluctuations may be relevant at the quantum phase transition.

Pure chromium deviates from the weak-coupling ground state for $P > 7.3$ GPa, at which pressure $\Delta < 19$ meV and $\xi > 125$ Å. By 9.5 GPa, the respective values are 9 meV and 260 Å. The resolution-limited CDW line scan (Fig. 3c) rules out both Fermi surface distortion and longitudinal fluctuations as candidates to disrupt the ground state on this length scale. We therefore posit that fluctuations transverse to $\mathbf{Q}$ on a length scale of several hundred angstroms render the weak-coupling theory inappropriate for $P > 7.3$ GPa. We do in fact observe line-shape broadening in the transverse direction consistent with such fluctuations, but cannot make a definitive claim given the crystal-lattice mosaic of individual domains. Similar anisotropic correlations are often found in three-dimensional systems such as quantum critical $CeCu_{6-x}Au_x$ (ref. 23) and CDW systems with quasi-one-dimensional electronic structure[24,25], and the probability of anisotropic fluctuations in chromium speaks to the quasi-one-dimensional dispersion relation of carriers at the highly nested Fermi surfaces. At the high pressures of our experiment, the SDW remains in the transverse phase for all T (ref. 14), leading to the possibility of transverse fluctuations of spin-polarization (S-domain[10]) correlation volumes.

This work addresses a question that is germane to the study of ordered electrons in general, namely that of how a single ground state emerges at a quantum phase transition in a system with multiple energy scales[26]. It is understood that, despite the weak-coupling density-wave ground state, chromium possesses a hierarchy of energy scales that bridge the traditional boundary between weak and strong coupling[15]. Here we show that the feature responsible for this strong-coupling phenomenology, namely the highly nested Fermi surfaces, survives into the quantum critical regime. It is therefore plausible that bosonic electron–hole pairs remain viable after the long-range phase coherence is destroyed. This presents the possibility that phenomena more typically associated with strongly coupled systems, such as local phase segregation or condensation of bosonic pairs, may be possible high-pressure ground states in this simplest of model systems. The crossover to a Bose–Einstein condensate of short-coherence-length pairs, observed in both cold atoms[3–5] and dimerized magnetic insulators[27], might be an organizing principle for a wider range of systems.

METHODS SUMMARY

We performed direct measurement of the CDW order parameter using monochromatic X-ray diffraction at insertion device beamline 4-ID-D of the Advanced Photon Source, Argonne National Laboratory. A silicon double-bounce monochromator was used to provide 20,000-keV X-rays, and a pair of palladium mirrors rejected higher harmonics and focused the beam to maximize the flux incident on a small sample volume. Pressure was generated using a home-built, helium-membrane-controlled diamond anvil cell that allowed the sample pressure to be changed *in situ* at base temperature with better than 0.05-GPa resolution. The pressure medium was a 4:1 methanol:ethanol solution. We determined the pressure *in situ* by measuring the lattice constant of a polycrystalline silver grain included in the pressure chamber, using the Birch equation of the first kind and a low-temperature bulk modulus of 108.96 GPa (ref. 15).

The pressure cell was mounted on an X-ray compatible closed-cycle cryostat, which was in turn mounted on a precision x–y–z translation stage on the sample stage of a psi-circle diffractometer. All data presented here were measured below 8 K; for $P > 7$ GPa, all data were measured below 6 K. Our samples are miniature single crystals with typical dimensions of $100 \times 100 \times 40\ \mu m^3$ and mosaic FWHM of 0.08° at the highest measured pressure. The strict sample requirements and preparation procedures are described elsewhere[15,28]. With the focused monochromatic X-ray beam, the highly collimated diffractometer and the synchrotron flux available at beamline 4-ID-D, we achieved a sensitivity of 5×10^{-10} relative to the lattice Bragg intensity (signal one-tenth of background), which is sufficient for tracking the CDW into the quantum critical regime. For comparison of disparate CDW Bragg peaks (for example, $I(4-2Q,1,\bar{2})/I(1,1,\bar{2})$ and $I(\bar{1},3-2Q,0)/I(\bar{1},1,0)$), we converted all intensities into units of $I_{CDW} \equiv I(2Q,0,0)/I(2,0,0)$ using a procedure outlined in ref. 15; reported values of I_{CDW} include contributions from all three domain types.

Received 19 December 2008; accepted 17 March 2009.

1. de la Cruz, C. *et al.* Magnetic order close to superconductivity in the iron-based layered $LaO_{1-x}F_xFeAs$ systems. *Nature* **453,** 899–902 (2008).
2. Morosan, E. *et al.* Superconductivity in Cu_xTiSe_2. *Nature Phys.* **2,** 544–550 (2006).
3. Regal, C. A., Greiner, M. & Jin, D. S. Observation of resonance condensation of fermionic atom pairs. *Phys. Rev. Lett.* **92,** 040403 (2004).
4. Chin, C. *et al.* Observation of the pairing gap in a strongly interacting Fermi gas. *Science* **305,** 1128–1130 (2004).
5. Chen, Q., Stajic, J., Tan, S. & Levin, K. BCS-BEC crossover: from high temperature superconductors to ultracold superfluids. *Phys. Rep.* **412,** 1–88 (2005).
6. Sachdev, S. Quantum criticality: competing ground states in low dimensions. *Science* **288,** 475–480 (2000).
7. Coleman, P. & Schofield, A. J. Quantum criticality. *Nature* **433,** 226–229 (2005).
8. Nagaosa, N. Superconductivity and antiferromagnetism in high-T_C cuprates. *Science* **275,** 1078–1079 (1997).
9. Broun, D. M. What lies beneath the dome? *Nature Phys.* **4,** 170–172 (2008).
10. Fawcett, E. Spin-density-wave antiferromagnetism in chromium. *Rev. Mod. Phys.* **60,** 209–283 (1988).
11. McWhan, D. B. & Rice, T. M. Pressure dependence of itinerant antiferromagnetism in chromium. *Phys. Rev. Lett.* **19,** 846–849 (1967).
12. Yeh, A. *et al.* Quantum phase transition in a common metal. *Nature* **419,** 459–462 (2002).
13. Lee, M., Husmann, A., Rosenbaum, T. F. & Aeppli, G. High resolution study of magnetic ordering at absolute zero. *Phys. Rev. Lett.* **92,** 187201 (2004).
14. Feng, Y. *et al.* Pressure-tuned spin and charge ordering in an itinerant antiferromagnet. *Phys. Rev. Lett.* **99,** 137201 (2007).
15. Jaramillo, R. *et al.* Chromium at high pressures: weak coupling and strong fluctuations in an itinerant antiferromagnet. *Phys. Rev. B* **77,** 184418 (2008).
16. Overhauser, A. W. Spin density waves in an electron gas. *Phys. Rev.* **128,** 1437–1452 (1962).
17. Young, C. Y. & Sokoloff, J. B. The role of harmonics in the first order antiferromagnetic to paramagnetic transition in chromium. *J. Phys. F* **4,** 1304–1319 (1974).
18. Hill, J. P., Helgesen, G. & Gibbs, D. X-ray-scattering study of charge- and spin-density waves in chromium. *Phys. Rev. B* **51,** 10336–10344 (1995).
19. Pfleiderer, C. Why first order quantum phase transitions are interesting. *J. Phys. Condens. Matter* **17,** S987–S997 (2005).

20. Werner, S. A., Arrott, A. & Kendrick, H. Temperature and magnetic-field dependence of the antiferromagnetism in pure chromium. *Phys. Rev.* **155,** 528–539 (1967).
21. Rice, T. M. Band-structure effects in itinerant antiferromagnetism. *Phys. Rev. B* **2,** 3619–3630 (1970).
22. Sokolov, D. A. *et al.* Elastic neutron scattering in quantum critical antiferromagnet $Cr_{0.963}V_{0.037}$. *Physica B (Amsterdam)* **403,** 1276–1278 (2008).
23. Schroder, A., Aeppli, G., Bucher, E., Ramazashvili, R. & Coleman, P. Scaling of magnetic fluctuations near a quantum phase transition. *Phys. Rev. Lett.* **80,** 5623–5626 (1998).
24. Sweetland, E., Tsai, C.-Y., Wintner, B. A., Brock, J. D. & Thorne, R. E. Measurement of the charge-density-wave correlation length in $NbSe_3$ by high-resolution X-ray scattering. *Phys. Rev. Lett.* **65,** 3165–3168 (1990).
25. DeLand, S. M., Mozurkewich, G. & Chapman, L. D. X-ray investigation of charge-density-wave pinning in blue bronze. *Phys. Rev. Lett.* **66,** 2026–2029 (1991).
26. Gegenwart, P. *et al.* Multiple energy scales at a quantum critical point. *Science* **315,** 969–971 (2007).
27. Giamarchi, T., Rüegg, C. & Tchernyshoyov, O. Bose–Einstein condensation in magnetic insulators. *Nature Phys.* **4,** 198–204 (2008).
28. Feng, Y. *et al.* Energy dispersive X-ray diffraction of charge density waves via chemical filtering. *Rev. Sci. Instrum.* **76,** 063913 (2005).
29. Strempfer, J. *et al.* Form-factor measurements on chromium with high-energy synchrotron radiation. *Eur. Phys. J. B* **14,** 63–72 (2000).
30. Rice, T. M., Barker, A. S., Halperin, B. I. & McWhan, D. B. Antiferromagnetism in chromium and its alloys. *J. Appl. Phys.* **40,** 1337–1343 (1969).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to J. Pluth for assistance with sample preparation, V. Prakapenka and GeoSoilEnviroCARS (Advanced Photon Source (APS), Argonne National Laboratory) for technical support and G. Aeppli for many discussions. The work at the University of Chicago was supported by the US National Science Foundation (NSF) Division of Materials Research. GeoSoilEnviroCARS is supported by the US NSF Earth Sciences and Department of Energy (DOE) Geosciences. Use of APS is supported by the US DOE Office of Basic Energy Sciences.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.F.R. (tfr@uchicago.edu).

Vol 459|21 May 2009|doi:10.1038/nature08053

LETTERS

Five-dimensional optical recording mediated by surface plasmons in gold nanorods

Peter Zijlstra[1], James W. M. Chon[1] & Min Gu[1]

Multiplexed optical recording provides an unparalleled approach to increasing the information density beyond 10^{12} bits per cm^3 (1 Tbit cm^{-3}) by storing multiple, individually addressable patterns within the same recording volume. Although wavelength[1–3], polarization[4–8] and spatial dimensions[9–13] have all been exploited for multiplexing, these approaches have never been integrated into a single technique that could ultimately increase the information capacity by orders of magnitude. The major hurdle is the lack of a suitable recording medium that is extremely selective in the domains of wavelength and polarization and in the three spatial domains, so as to provide orthogonality in all five dimensions. Here we show true five-dimensional optical recording by exploiting the unique properties of the longitudinal surface plasmon resonance (SPR) of gold nanorods. The longitudinal SPR exhibits an excellent wavelength and polarization sensitivity, whereas the distinct energy threshold required for the photothermal recording mechanism provides the axial selectivity. The recordings were detected using longitudinal SPR-mediated two-photon luminescence, which we demonstrate to possess an enhanced wavelength and angular selectivity compared to conventional linear detection mechanisms. Combined with the high cross-section of two-photon luminescence, this enabled non-destructive, crosstalk-free readout. This technique can be immediately applied to optical patterning, encryption and data storage, where higher data densities are pursued.

The concept of five-dimensional patterning is illustrated in Fig. 1. The sample consists of a multilayered stack in which thin recording layers (~1 μm) are separated by a transparent spacer (~10 μm). In both the wavelength and polarization domains, three-state multiplexing is illustrated to provide a total of nine multiplexed states in one recording layer. The key to successfully realizing such five-dimensional encoding is a recording material that (1) is orthogonal in all dimensions, in both recording and readout, (2) is able to provide multiple recording channels in each dimension, and (3) is stable in ambient conditions and can be read out non-destructively. Existing multiplexing techniques[1–8] are only orthogonal in one dimension (either wavelength or polarization), and often ambient conditions and readout degrades the recorded patterns through unwanted isomerization or photobleaching.

We show that a recording material based on plasmonic gold nanorods meets all the above criteria. Gold nanorods have been extensively used in a wide range of applications because of their unique optical and photothermal properties. The narrow longitudinal SPR linewidth of a gold nanorod (100–150 meV, or ~45–65 nm in the near-infrared[14,15]; Supplementary Fig. 1), combined with the dipolar optical response, allows us to optically address only a small subpopulation of nanorods in the laser irradiated region. We use this selectivity to achieve longitudinal SPR mediated recording and readout governed by photothermal reshaping and two-photon luminescence (TPL) detection, respectively.

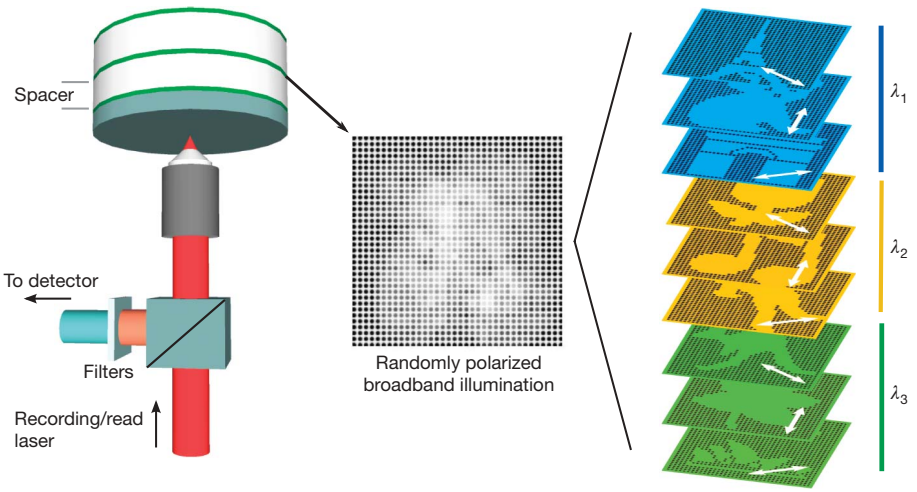


Figure 1 | Sample structure and patterning. Left, the sample consists of thin recording layers of spin-coated polyvinyl alcohol doped with gold nanorods, on a glass substrate. These recording layers were spaced by a transparent pressure-sensitive adhesive with a thickness of 10 μm. In the recording layers, we patterned multiple images using different wavelengths (λ_{1-3}) and polarizations of the recording laser. Middle, when illuminated with unpolarized broadband illumination, a convolution of all patterns will be observed on the detector (filters attenuate the reflected readout laser light). Right, when the right polarization and wavelength is chosen, the patterns can be read out individually without crosstalk.

[1]Centre for Micro-Photonics, Faculty of Engineering and Industrial Sciences, Swinburne University of Technology, PO Box 218, Hawthorn, Victoria 3122, Australia.

During recording, the absorption of a laser pulse induces a temperature rise in the selected nanorods. For sufficiently high laser pulse energy, the selected nanorods will heat up to above the threshold melting temperature, and transform their shape into shorter rods or spherical particles[16,17]. This results in a depleted population of nanorods with a certain aspect ratio and orientation (Fig. 2a), and hence a polarization dependent bleaching occurs in the extinction profile (Supplementary Fig. 2). Despite the single photon excitation of the longitudinal SPR, the threshold of the photothermal melting confines the writing process axially to within the focal volume and provides the ability to record three-dimensionally. This is in stark contrast to single photon recording by photobleaching or photoisomerization, where the out-of-focus laser light would still induce recording.

In Fig. 2b we show the transmission electron microscope images and the extinction profiles of the three distributions of gold nanorods that we used in the recording layers. To confirm the selective reshaping, a mixture of the gold nanorods depicted in Fig. 2b was spin-coated on a glass coverslip. We acquired scanning electron microscope (SEM) images before (Fig. 2c) and after (Fig. 2d) irradiation with a single, linearly polarized femtosecond laser pulse ($\lambda = 840$ nm and pulse energy 0.28 nJ in the focal plane of the objective). We find that only nanorods with an aspect ratio of 3.4 ± 0.9 within an angular range of 25° with respect to the horizontal laser light polarization were affected by the laser pulse (averaged over 20 reshaped particles). Some of the rods were propelled from the glass interface owing to a rapid change in the centre of mass during the melting process, which was observed previously for gold triangles[18]. Such lift-off is prevented in our recording medium, as the nanorods are embedded in a thick polymer layer (Supplementary Fig. 3).

We imaged the recordings using longitudinal SPR mediated TPL. Previous reports on patterning in gold nanoparticles all used linear detection processes based on scattering[2] or extinction[5–7,19,20]. However, a nonlinear detection process such as TPL has a significantly higher angular and wavelength sensitivity. To demonstrate this, we acquired the scattering and TPL excitation profiles of a single gold nanorod (average aspect ratio 3, average size 90×30 nm) as a function of both wavelength and polarization (Fig. 3a and b). The nonlinear character of the TPL induces an excitation profile linewidth that is almost 60% narrower than the linewidth of the linear scattering spectrum (Fig. 3a, Supplementary Fig. 4). Furthermore, we find a reduction of almost 50% in the width of the angular excitation profile (Fig. 3b), which is in good agreement with previous observations[21,22]. The observed narrowing of the spectral and angular excitation profiles significantly reduces interference in the readout between neighbouring recording channels. Even more so, the axial sectioning induced by two-photon excitation allows for crosstalk-free readout of closely spaced layers. The most fascinating property of TPL is that it is most efficiently excited on resonance with the linear plasmon absorption band[21,22], enabling single photon recording and multi-photon readout using one and the same wavelength.

The TPL brightness of gold nanorods was characterized by calculating their TPL action cross-section ($=\eta\sigma_2$, where η is the luminescence quantum yield and σ_2 is the two-photon absorption cross-section) of a single gold nanorod following the method of ref. 23. From a TPL raster scan of isolated gold nanorods (average aspect ratio 4, average size 44×12 nm), we estimate the TPL action cross-section to be $\sim 3 \times 10^4$ GM (Göppert-Mayer) for excitation on resonance with the longitudinal SPR (for more details, see Supplementary Information page 6). From the TPL action cross-section, σ_2 can be calculated if η is known. Previous reports on photoluminescence of gold nanoparticles[24–26] suggest that the quantum yield of the nanorod geometry is drastically increased compared to spherical particles[25] or films[27]. The so-called 'lightning rod effect' around a nanorod is known to enhance the local field strength and the radiative decay rate via coupling to the SPR, and has been used to explain the observed increase in quantum yield from 10^{-10} for a film to $\sim 10^{-4}$ for rods with a size and aspect ratio similar to the ones studied here[24]. Assuming the reported quantum yield, we estimate that σ_2 is $\sim 3 \times 10^8$ GM. This is to our knowledge the highest σ_2 ever observed, with the previous highest report being 3.5×10^6 GM for a 4-nm gold nanoparticle[26]. We propose that this drastic increase is caused by the two orders of magnitude larger volume and intrinsically higher optical cross-section of our nanorods compared to the 4-nm-diameter gold nanoparticles used in ref. 26. The current direct measurement of the TPL action cross-section is also one order of magnitude larger than the value of 2.3×10^3 GM reported in ref. 21, which was indirectly determined by comparing the brightness of a single rhodamine 6G

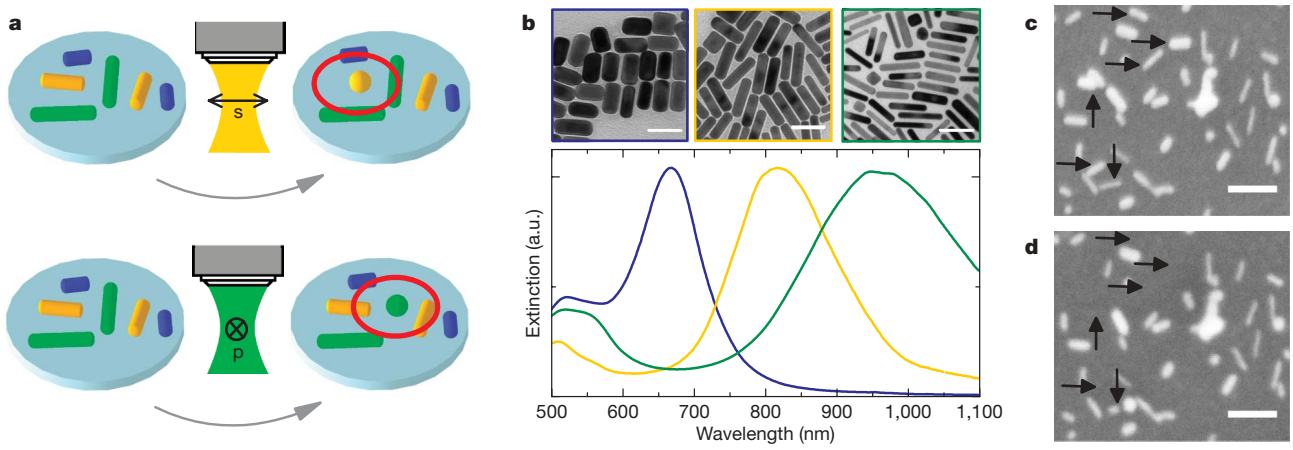


Figure 2 | Photothermal patterning. **a**, A schematic illustration of the patterning mechanism. The patterning is governed by photothermal reshaping of the gold nanorods in the focal volume of the focusing objective; this reshaping is selective in terms of aspect ratio and orientation. A linear polarized laser pulse will only be absorbed by gold nanorods that are aligned to the laser light polarization and which exhibit an absorption cross-section that matches the laser wavelength. Top, s-polarized laser light with a wavelength of 840 nm will only affect the nanorods with an intermediate aspect ratio that are aligned to the laser polarization (the reshaped rod is indicated). Bottom, p-polarized laser light with a wavelength of 980 nm only reshapes the high-aspect-ratio gold nanorods aligned to the laser light polarization (reshaped rod is indicated). **b**, Normalized extinction spectra of the as-prepared gold nanorod solutions. a.u., arbitrary units. Insets show transmission electron micrographs of the gold nanorods on a copper grid. The average sizes of the nanorods are, from left to right, 37×19 nm (aspect ratio 2 ± 1), 50×12 nm (aspect ratio 4.2 ± 1), and 50×8 nm (aspect ratio 6 ± 2). Scale bars, 50 nm. Each recording layer in the multi-layered sample is doped with a mixture of these gold nanorods to form an inhomogeneously broadened extinction profile. **c**, **d**, SEM images of gold nanorods spin-coated on an indium tin oxide coated glass coverslip before (**c**) and after (**d**) irradiation with a single femtosecond laser pulse of 840 nm with horizontal polarization. Rods affected by the laser pulse are arrowed. Scale bars, 100 nm.

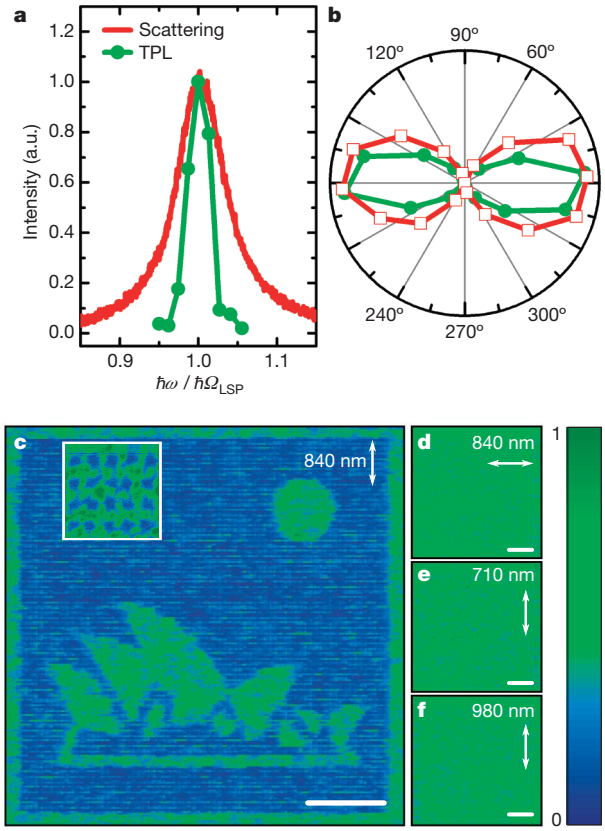


Figure 3 | Readout using two-photon luminescence. **a**, Normalized white light scattering spectrum (red solid line) and two-photon luminescence (TPL) excitation profile (green circles) as a function of photon energy $\hbar\omega$ for individual gold nanorods with approximate dimensions of 90 × 30 nm. The profiles were centred around the longitudinal SPR energy $\hbar\Omega_{LSP}$. **b**, Polarized scattering (open squares) and TPL intensity (filled circles) versus the polarization of the excitation light. **c**, Normalized TPL raster scan of an image patterned using single laser pulses per pixel with a wavelength of 840 nm and vertical polarization. (Panels **c**–**f** are labelled with laser wavelength and a double-headed arrow indicating polarization direction.) The TPL was excited with the same wavelength and polarization as was used for the patterning. The pattern is 75 × 75 pixels, with a pixel spacing of 1.33 μm. Inset, high magnification image of the recording (size 7 × 7 μm). **d**–**f**, Images obtained when the TPL is excited at 840 nm with horizontal polarization (**d**), at 710 nm with vertical polarization (**e**), and at 980 nm with vertical polarization (**f**). Scale bars, 20 μm.

molecule to a single gold nanorod. The large TPL action cross-section allows us to non-destructively image the recordings by using very low excitation powers.

To illustrate the readout using TPL, we first recorded a single image using vertically polarized laser light with a wavelength of 840 nm (Fig. 3c). Pixels were written using a single femtosecond laser pulse focused through a 0.95 NA objective lens. Using single laser pulses per pixel prevents adverse accumulative thermal effects on the matrix, which were observed when a high repetition rate pulse train was employed[20]. The recorded pattern was retrieved by raster scanning the sample and detecting the TPL signal from the gold nanorods. The TPL was excited with the same wavelength and polarization that was used for the patterning. The pixels exhibit a lower TPL signal owing to a depleted population of nanorods with a longitudinal SPR on resonance with the readout laser light. After deconvolution with the response function of the imaging objective, we find an average pixel size of 500 ± 100 nm, which is in good agreement with the expected diffraction limit of 470 nm. Because all the rods on resonance with the laser light were reshaped, the contrast of the pixels (defined as $|I_{pixel} - I_{bg}|/(I_{pixel} + I_{bg})$, with I_{pixel} and I_{bg} the pixel and background TPL signal, respectively) was 1. When this image was read in transmission mode, the contrast was found to be 0.05. Considering the nanorod concentration (Methods), we estimate that this contrast arises from the reshaping of ~30 nanorods in the focal volume. We do not observe any contrast when the TPL is excited with a horizontally polarized laser beam (Fig. 3d) or when the wavelength is tuned to 700 nm or 980 nm (Fig. 3e and f). This indicates that only the subpopulation of nanorods with a longitudinal SPR on resonance with the recording laser light has reshaped.

Using the longitudinal SPR mediated recording and readout mechanisms, we achieved for the first time five-dimensional optical recording. In Fig. 4 we show TPL raster scans of 18 images, all patterned in the same area. The patterning was conducted using a single femtosecond laser pulse per pixel at wavelengths of 700 nm, 840 nm and

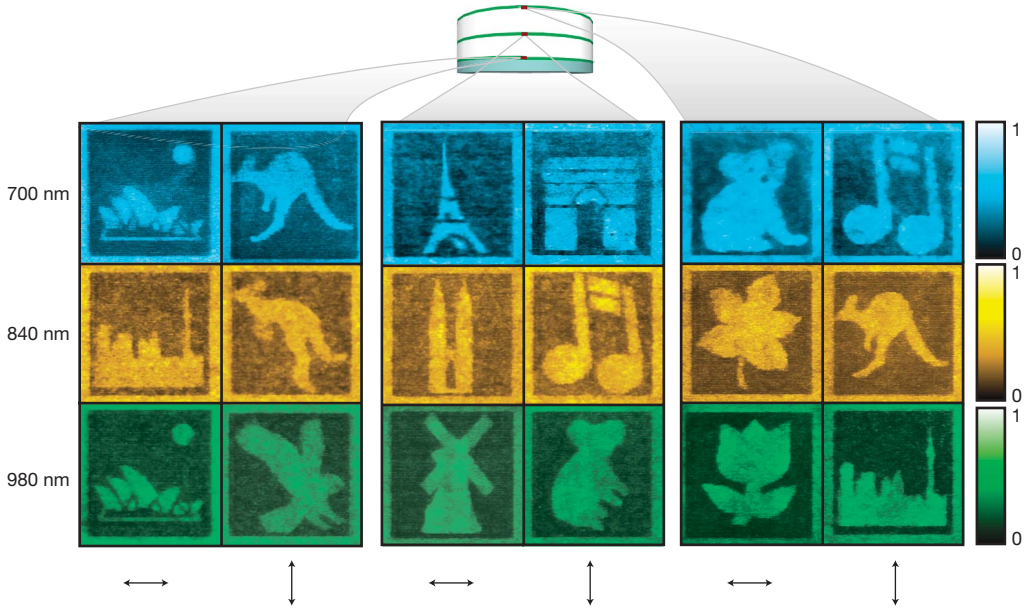


Figure 4 | Five-dimensional patterning and readout. Normalized TPL raster scans of 18 patterns encoded in the same area using two laser light polarizations and three different laser wavelengths. Patterns were written in three layers spaced by 10 μm. The recording laser pulse properties are indicated (wavelength at left, polarization at bottom). The recordings were retrieved by detecting the TPL excited with the same wavelength and polarization as employed for the recording. The size of all images is 100 × 100 μm, and the patterns are 75 × 75 pixels.

980 nm, with both horizontal and vertical polarization (see Methods for more details). Images were patterned in three layers, with a layer spacing of 10 μm and a bit spacing of 1.33 μm. The laser pulse energy and wavelength used for patterning were optimized to minimize cross-talk between the different recording channels. Although we used a femtosecond pulse laser for patterning, the recording can also be performed with a continuous wave laser or laser diode, paving the way for a low-cost recording apparatus (Supplementary Fig. 5).

Using this technique, improved security imprinting and encryption can be realized, where the added dimensions can act as an extended and counterfeit-proof encryption key. In such applications, it would be highly beneficial to have immediate access to the patterns without the need for raster scanning. In order to test this, we have recorded polarization multiplexed images in three layers spaced by 40 μm, which were then read out using a charge coupled device (CCD) and a white light source (Supplementary Movie 1). This technique allows for one-shot readout of patterns, and when multiple CCDs are used it can provide instant and simultaneous readout of all recorded patterns. Additionally, the modulation of the transmission introduced by the patterns can be used as a polarization- and wavelength-dependent signal modulator for optical devices. For example, the large extinction of gold nanorods allows for customized modulation in multiple wavelength bands of a supercontinuum light source. This can be accomplished in a single (on chip) filter, which does not suffer from bleaching and exhibits a high damage threshold of $>10\ \mathrm{mJ\ cm^{-2}}$ (based on the threshold energy required for recording a pixel).

Most importantly, the presented technique could be highly beneficial for high density optical data storage. As demonstrated in Fig. 4, incorporation of two polarization and three wavelength channels, a 10-μm spacer layer, and a bit spacing equal to the bit diameter of 0.75 μm amounts to a bit density of 1.1 $\mathrm{Tbit\ cm^{-3}}$. This results in a disk capacity of 1.6 Tbyte for a DVD sized disk. We have performed recording and readout in ten layers (Supplementary Fig. 6), demonstrating the feasibility of our technology for application on a disk. Further improvement in data capacity is warranted if three-state polarization encoding (Supplementary Fig. 7) is combined with a thinner spacer. A threefold reduction in spacer layer thickness is feasible, considering the recording layer thickness of 1 μm, resulting in a disk capacity of 7.2 Tbyte. Our bit-by-bit recording technique is fully compatible with existing drive technology, and allows for recording speeds up to 1 $\mathrm{Gbit\ s^{-1}}$ when a high repetition rate laser source is used[28]. Owing to the low recording pulse energy (<0.5 nJ per pulse), a drastic increase in recording speed can be achieved when a supercontinuum light source is used for simultaneous recording in all channels.

METHODS SUMMARY

Sample preparation. Gold nanorods with average aspect ratios of 2.3 ± 1, 4.3 ± 1 and 6 ± 2 were prepared using wet chemical synthesis[29,30]. The nanorod solutions were then combined to obtain a 'flat' extinction profile in the 700–1,000 nm wavelength range. The nanorods were mixed with a 15 wt% polyvinyl alcohol solution, and spin-coated on a glass coverslip. The thickness of this layer was 1 ± 0.2 μm, measured using an atomic force microscope. The approximate nanorod concentration in the film was 400 ± 50 nM (~200 nanorods in the focal volume of the 0.95 NA objective lens). Subsequently, a transparent pressure-sensitive adhesive (LINTEC Co.) with a thickness of 10 ± 1 μm and a refractive index of 1.506 was laminated onto the spin-coated layer. This process was repeated until the desired number of layers was reached (for more details, see Supplementary Information).
Optical set-up. Both recording and readout were conducted in the same home-built microscope. For recording, an electro-optic modulator selected single pulses from the pulse train of a femtosecond pulse laser (SpectraPhysics Tsunami, 100 fs pulse duration, 82 MHz repetition rate, tunable between 690 and 1,010 nm). The laser pulses were focused onto the sample through a high NA objective lens (Olympus 0.95 NA 40×, coverslip corrected). For readout, the TPL of the nanorods was excited using the 82 MHz output from the femtosecond laser. The TPL signal was directed to a photomultiplier tube (Hamamatsu H7422P40) and was detected in the 400–600 nm wavelength range. To prevent erasure of the patterns, the pulse energy of the readout laser was almost three orders of magnitude lower than the patterning pulse energy (for more details see Supplementary Information).

Received 27 November 2008; accepted 2 April 2009.

1. Moerner, W. E. *Persistent Spectral Hole-Burning: Science and Applications* (Springer, 1988).
2. Ditlbacher, H., Krenn, J. R., Lamprecht, B., Leitner, A. & Aussenegg, F. R. Spectrally coded optical data storage by metal nanoparticles. *Opt. Lett.* **25,** 563–565 (2000).
3. Pham, H. H., Gourevich, I., Oh, J. K., Jonkman, J. E. N. & Kumacheva, E. A multidye nanostructured material for optical data storage and security data encryption. *Adv. Mater.* **16,** 516–520 (2004).
4. Alasfar, S. *et al.* Polarization-multiplexed optical memory with urethane-urea copolymers. *Appl. Opt.* **38,** 6201–6204 (1999).
5. Niidome, Y., Urakawa, S., Kawahara, M. & Yamada, S. Dichroism of poly(vinylalcohol) films containing gold nanorods induced by polarized pulsed-laser irradiation. *Jpn J. Appl. Phys.* **42,** 1749–1750 (2003).
6. Wilson, O., Wilson, G. J. & Mulvaney, P. Laser writing in polarized silver nanorod films. *Adv. Mater.* **14,** 1000–1004 (2002).
7. Pérez-Juste, J., Rodríguez-González, B., Mulvaney, P. & Liz-Marzán, L. M. Optical control and patterning of gold-nanorod-poly(vinyl alcohol) nanocomposite films. *Adv. Funct. Mater.* **15,** 1065–1071 (2005).
8. Li, X. P., Chon, J. W. M., Wu, S. H., Evans, R. A. & Gu, M. Rewritable polarization-encoded multilayer data storage in 2,5-dimethyl-4-(p-nitrophenylazo)anisole doped polymer. *Opt. Lett.* **32,** 277–279 (2007).
9. Strickler, J. & Webb, W. Three-dimensional optical data storage in refractive media by two-photon point excitation. *Opt. Lett.* **16,** 1780–1782 (1991).
10. Heanue, J. F., Bashaw, M. C. & Hesselink, L. Volume holographic storage and retrieval of digital data. *Science* **265,** 749–752 (1994).
11. Cumpston, B. H. *et al.* Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* **398,** 51–54 (1999).
12. Kawata, S. & Kawata, Y. Three-dimensional optical data storage using photochromic materials. *Chem. Rev.* **100,** 1777–1788 (2000).
13. Day, D., Gu, M. & Smallridge, A. Rewritable 3D bit optical data storage in a PMMA-based photorefractive polymer. *Adv. Mater.* **13,** 1005–1007 (2001).
14. Novo, C. *et al.* Contributions from radiation damping and surface scattering to the linewidth of the longitudinal plasmon band of gold nanorods: a single particle study. *Phys. Chem. Chem. Phys.* **8,** 3540–3546 (2006).
15. Sönnichsen, C. *et al.* Drastic reduction of plasmon damping in gold nanorods. *Phys. Rev. Lett.* **88,** 077402 (2002).
16. Chang, S. S., Shih, C. W., Chen, C. D., Lai, W. C. & Wang, C. R. C. The shape transition of gold nanorods. *Langmuir* **15,** 701–709 (1999).
17. Link, S., Burda, C., Nikoobakht, B. & El-Sayed, M. A. Laser-induced shape changes of colloidal gold nanorods using femtosecond and nanosecond laser pulses. *J. Phys. Chem. B* **104,** 6152–6163 (2000).
18. Habenicht, A., Olapinski, M., Burmeister, F., Leiderer, P. & Boneberg, J. Jumping nanodroplets. *Science* **309,** 2043–2045 (2005).
19. Chon, J. W. M., Bullen, C., Zijlstra, P. & Gu, M. Spectral encoding on gold nanorods doped in a silica sol-gel matrix and its application to high density optical data storage. *Adv. Funct. Mater.* **17,** 875–880 (2007).
20. Zijlstra, P., Chon, J. W. M. & Gu, M. Effect of heat accumulation on the dynamic range of a gold nanorod doped nanocomposite for optical laser writing and patterning. *Opt. Express* **15,** 12151–12160 (2007).
21. Wang, H. F. *et al. In vitro* and *in vivo* two-photon luminescence imaging of single gold nanorods. *Proc. Natl Acad. Sci. USA* **102,** 15752–15756 (2005).
22. Bouhelier, A. *et al.* Surface plasmon characteristics of tunable photoluminescence in single gold nanorods. *Phys. Rev. Lett.* **95,** 267405 (2005).
23. Xu, C. & Webb, W. W. Measurement of two-photon excitation cross sections of molecular fluorophores with data from 690 to 1050 nm. *J. Opt. Soc. Am. B* **13,** 481–491 (1996).
24. Mohamed, M. B., Volkov, V., Link, S. & El-Sayed, M. A. The 'lightning' gold nanorods: fluorescence enhancement of over a million compared to the gold metal. *Chem. Phys. Lett.* **317,** 517–523 (2000).
25. Dulkeith, E. *et al.* Plasmon emission in photoexcited gold nanoparticles. *Phys. Rev. B* **70,** 205424 (2004).
26. Ramakrishna, G., Varnavski, O., Kim, J., Lee, D. & Goodson, T. Quantum sized gold clusters as efficient two photon absorbers. *J. Am. Chem. Soc.* **130,** 5032–5033 (2008).
27. Mooradian, A. Photoluminescence of metals. *Phys. Rev. Lett.* **22,** 185–187 (1969).
28. Tanaka, T. & Kawata, S. Three-dimensional multi-layered fluorescent optical disk. In *Technical Digest Int. Symp. Opt. Mem.* Tu-G-01 (Adthree Publishing, Tokyo, 2007).
29. Nikoobakht, B. & El-Sayed, M. A. Preparation and growth mechanism of gold nanorods (NRs) using seed-mediated growth method. *Chem. Mater.* **15,** 1957–1962 (2003).
30. Zijlstra, P., Bullen, C., Chon, J. W. M. & Gu, M. High-temperature seedless synthesis of gold nanorods. *J. Phys. Chem. B* **110,** 19315–19318 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the LINTEC Corporation for supplying the pressure-sensitive adhesive and the Australian Research Council for financial support. We thank R. Evans, W. Rowlands and D. Buso for carefully reading the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.W.M.C. (JChon@groupwise.swin.edu.au).

Vol 459|21 May 2009|doi:10.1038/nature08016

LETTERS

Self-assembly of DNA into nanoscale three-dimensional shapes

Shawn M. Douglas[1,2,3], Hendrik Dietz[1,2], Tim Liedl[1,2], Björn Högberg[1,2], Franziska Graf[1,2,3] & William M. Shih[1,2,3]

Molecular self-assembly offers a 'bottom-up' route to fabrication with subnanometre precision of complex structures from simple components[1]. DNA has proved to be a versatile building block[2–5] for programmable construction of such objects, including two-dimensional crystals[6], nanotubes[7–11], and three-dimensional wireframe nanopolyhedra[12–17]. Templated self-assembly of DNA[18] into custom two-dimensional shapes on the megadalton scale has been demonstrated previously with a multiple-kilobase 'scaffold strand' that is folded into a flat array of antiparallel helices by interactions with hundreds of oligonucleotide 'staple strands'[19,20]. Here we extend this method to building custom three-dimensional shapes formed as pleated layers of helices constrained to a honeycomb lattice. We demonstrate the design and assembly of nanostructures approximating six shapes—monolith, square nut, railed bridge, genie bottle, stacked cross, slotted cross—with precisely controlled dimensions ranging from 10 to 100 nm. We also show hierarchical assembly of structures such as homomultimeric linear tracks and heterotrimeric wireframe icosahedra. Proper assembly requires week-long folding times and calibrated monovalent and divalent cation concentrations. We anticipate that our strategy for self-assembling custom three-dimensional shapes will provide a general route to the manufacture of sophisticated devices bearing features on the nanometre scale.

The assembly of a target three-dimensional shape using the honeycomb-pleat-based strategy described here can be conceptualized as laying down the scaffold strand into an array of antiparallel helices (Fig. 1a) where helix $m + 1$ has a preferred attachment angle to helix m of $\pm 120^\circ$ degrees with respect to the attachment of helix $m - 1$ to helix m (Fig. 1b, c). This angle is determined by the relative register along the helical axes of the Holliday-junction crossovers that connect helix $m + 1$ to helix m versus those that connect helix $m - 1$ *to* helix m. Branching flaps are allowed as well (Supplementary Note S1).

The design procedure is analogous to sculpture from a porous crystalline block. Here the block is a honeycomb lattice of antiparallel scaffold helices (Fig. 1d). Complementary staple strands wind in an antiparallel direction around the scaffold strands to assemble B-form double helices that are assigned initial geometrical parameters (that can later be adjusted to account for interhelical repulsion) of 2.0 nm diameter, 0.34 nm per base-pair rise, and 34.3° per base-pair mean twist (or 21 base pairs every two turns). Crossovers between adjacent staple helices are restricted to intersections between the block and every third layer of a stack of planes orthogonal to the helical axes, spaced apart at intervals of seven base pairs or two-thirds of a turn (Fig. 1c). Crossovers between adjacent scaffold helices are permitted at positions displaced upstream or downstream of the corresponding staple-crossover points by five base pairs or a half-turn.

The first steps in the design process are carving away duplex segments from the block to define the target shape, and then introducing scaffold crossovers at a subset of allowed positions so as to create a singular scaffold path that visits all remaining duplex segments. Next, staple crossovers are added at all permitted positions on the shape that are not five base pairs away from a scaffold crossover; this exception maintains the local crossover density along any helix–helix interface at roughly one per 21 base pairs. Nicks are introduced into staple helices to define staple strands whose lengths are between 18 and 49 bases inclusive, with a mean between 30 and 42 bases. Sometimes staple crossovers are removed at the edges of the shapes to allow adjustment of staple lengths to preferred values. Unpaired scaffold bases are often introduced at the ends of helices to minimize undesired multimerization, or else to accommodate later addition of connecting staple strands that mediate desired multimerization. The final step is to thread the actual scaffold sequence on the target scaffold path to determine the Watson–Crick-complementary sequences of the staple strands.

Design steps and assignment of staple sequences for the shapes presented here were aided by manual rendering of strand diagrams in Adobe Illustrator and by writing *ad hoc* computer programs to produce staple sequences corresponding to those diagrams. This process was very time-consuming and error-prone even for trained DNA nanotechnologists. More recently, we have developed caDNAno, a graphical-interface-based computer-aided-design environment for assisting in honeycomb-pleated-origami design[21], and have ported all the objects described in this article into this framework (Supplementary Note S2). With caDNAno, an individual with no prior knowledge of programming or DNA structure can complete a short tutorial and then be capable of generating sequences within a day for building a new shape comparable in complexity to the examples demonstrated here.

As with flat DNA origami[19], assembly of three-dimensional, honeycomb-pleated DNA origami proceeds in a one-pot reaction, after rapid heating followed by slow cooling, between a scaffold strand and the hundreds of oligonucleotide staple strands that direct its folding into the target shape. Successful folding was observed for a panel of five structural targets (detailed schematics in Supplementary Note S2) each produced by mixing 10 nM scaffold strands derived from the single-stranded genome of the M13 bacteriophage (preparation described in Supplementary Note S1), 50 nM of every oligonucleotide staple strand, purified by reverse-phase cartridge (Bioneer Inc.), buffer and salts including 5 mM Tris + 1 mM EDTA (pH 7.9 at 20 °C), 16 mM $MgCl_2$, and subjecting the mixture to a thermal-annealing ramp that cooled from 80 °C to 60 °C over the course of 80 min and then cooled from 60 °C to 24 °C over the course of 173 h. Objects were electrophoresed on a 2% agarose gel containing 45 mM Tris borate + 1 mM EDTA (pH 8.3 at 20 °C), and 11 mM $MgCl_2$ at 70 V for four hours cooled by an ice-water bath, monomer bands were excised, DNA was recovered by physical extraction from the excised band, and the objects were imaged using transmission electron microscopy after negative-staining by uranyl formate. The

[1]Department of Cancer Biology, Dana-Farber Cancer Institute, [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts 02138, USA.

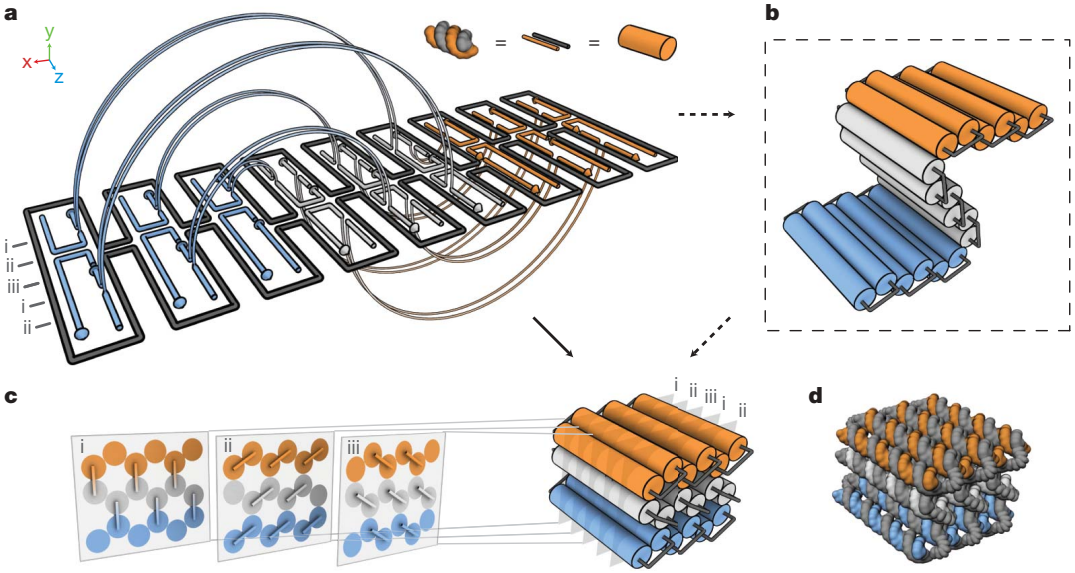


Figure 1 | Design of three-dimensional DNA origami. a, Double helices comprised of scaffold (grey) and staple strands (orange, white, blue) run parallel to the z-axis to form an unrolled two-dimensional schematic of the target shape. Phosphate linkages form crossovers between adjacent helices, with staple crossovers bridging different layers shown as semicircular arcs. **b**, Cylinder model of a half-rolled conceptual intermediate. Cylinders represent double helices, with loops of unpaired scaffold strand linking the ends of adjacent helices. **c**, Cylinder model of folded target shape. The honeycomb arrangement of parallel helices is shown in cross-sectional slices (i–iii) parallel to the *x*–*y* plane, spaced apart at seven base-pair intervals that repeat every 21 base pairs. All potential staple crossovers are shown for each cross-section. **d**, Atomistic DNA model of shape from **c**.

fraction of scaffold strands that were incorporated into monomeric species after folding varied from 7% to 44% for these targets as estimated by ethidium-bromide fluorescence intensity. Gel-purified particles were generally observed to be monodisperse with a homogenous shape (Fig. 2f); defect analysis for a series of related objects can be found elsewhere[21].

The five objects displayed in Fig. 2 demonstrate the generality of this honeycomb-pleated origami approach in approximating various three-dimensional shapes. Figure 2a shows a structure resembling a monolith, assembled in the form of a honeycomb-pleated block as in Fig. 1, except with ten layers instead of three. Particles display the predicted pattern of holes and stripes consistent with a honeycomb lattice of cylinders. Figure 2b shows a square nut, the cross-section of which is a block of the honeycomb lattice with an internal pore shaped like a six-pointed star. Figure 2c shows a structure that resembles a bridge with hand rails. This shape demonstrates that different cross-section patterns can be implemented along the helical axis. Figure 2d shows a slotted cross, a structure composed of two honeycomb-lattice-based domains that sit at 90° to one another. One domain is H-shaped, the other is O-shaped. The centre of the H-domain passes through the slot of the O-domain, and the two domains are connected by a pair of Holliday-junction crossovers derived from the scaffold strand. The 90° angle between domains is enforced by steric collisions between the ends of helices on the H-domain and the sides of helices on the O-domain. The fifth particle image for the slotted cross in Fig. 2d shows a defective particle, where the slot in the O-domain can be seen clearly. Figure 2e shows a stacked cross, where again two domains sit at 90° to one another. One domain is C-shaped, the other domain resembles a pod with a cavity. The pod domain consists of four sub-modules that are each connected to the C-shaped domain by a Holliday-junction crossover derived from the scaffold strand. Upon folding, the sub-modules connect to each other by staple linkages, enforcing a rotation to yield the complete pod domain oriented 90° to the C-module.

For the monolith, an effective diameter of 2.4 nm (±0.1 nm standard deviation, s.d.) per individual double helix was observed (Fig. 2g, h), while for the square nut an effective diameter of 2.1 nm (±0.1 nm s.d.) per individual double helix was observed (Fig. 2i, j). Assuming an unhydrated helical diameter of 2.0 nm (although the hydrodynamic helical diameter has been estimated[22] as 2.2–2.6 nm), this observation suggests the presence of inter-helical gaps produced by electrostatic repulsion[8] of the order of 0.1–0.4 nm, significantly less than the 1.0 nm gap size estimated for Rothemund flat origami. This discrepancy is probably related to the roughly twofold higher density of crossovers present in the honeycomb-pleated origami. Differences in effective helix diameter between architectures may originate in part from staining artefacts (for example, cavities where large amounts of positively charged stain accumulate, or flattening).

Three key determinants for folding of honeycomb-pleated origami were investigated: duration of thermal ramp, divalent-cation concentration, and monovalent-cation concentration. Folding with short thermal ramps (Fig. 3b, lefthand lanes), low concentrations of $MgCl_2$ (Fig. 3d, lefthand lanes), or high concentrations of NaCl (Fig. 3f, lefthand lanes) yielded a slowly migrating species upon agarose-gel electrophoresis and grossly misshapen objects as observed by transmission electron microscopy (for example, see Fig. 3c). In contrast, week-long thermal annealing at higher concentrations of $MgCl_2$ combined with low concentrations of NaCl yielded a fast-migrating species upon agarose-gel electrophoresis and well-folded particles as observed by electron microscopy (Fig. 3e), along with lower mobility bands corresponding to multimerized and aggregated objects. The apparent trend was that increasing agarose-gel mobility correlated with improvement of quality of folding as observed by transmission electron microscopy, suggesting that correctly folded structures tend to be more compact than misfolded versions.

Divalent cations thus appear to accelerate the rate of proper folding and increase the amount of undesired aggregation whereas monovalent cations appear to decelerate the rate of proper folding and decrease the amount of undesired aggregation. Many of the structures require week-long thermal ramps for proper folding, even under idealized divalent- and monovalent-cation concentrations. Divalent cations may accelerate target folding by specific stabilization of Holliday-junction crossovers[23] and by nonspecific stabilization of compact DNA[24] folding intermediates, although they may also stabilize nontarget aggregates by a similar mechanism. Monovalent-cation binding might compete with divalent-cation binding, and thereby antagonize both target compaction and nontarget aggregation, analogous to how such binding inhibits multivalent-cation-induced DNA condensation[25]. Folding

of simpler DNA-origami structures such as the six-helix-bundle nanotube is much more robust to variations in annealing conditions (Supplementary Note S1); the Rothemund flat origami and these simpler nanotube structures could be folded with 72 min ramps. Presumably, multilayered structures must traverse more difficult kinetic traps, perhaps owing in part to the larger density of crossovers, in part to issues of local folding and unfolding in the confined space between two or more layers of DNA helices, and in part to the difficulties in reaching a high density of DNA in the final folded object, similar to that found in high-pressure virus capsids[26].

One of the target shapes presented in Fig. 3 — the genie bottle (strand diagram in Supplementary Note S2) — was folded with two different scaffold sequences. Its full size takes up only 4,500 base pairs. One scaffold sequence used for folding was a modified M13 genome

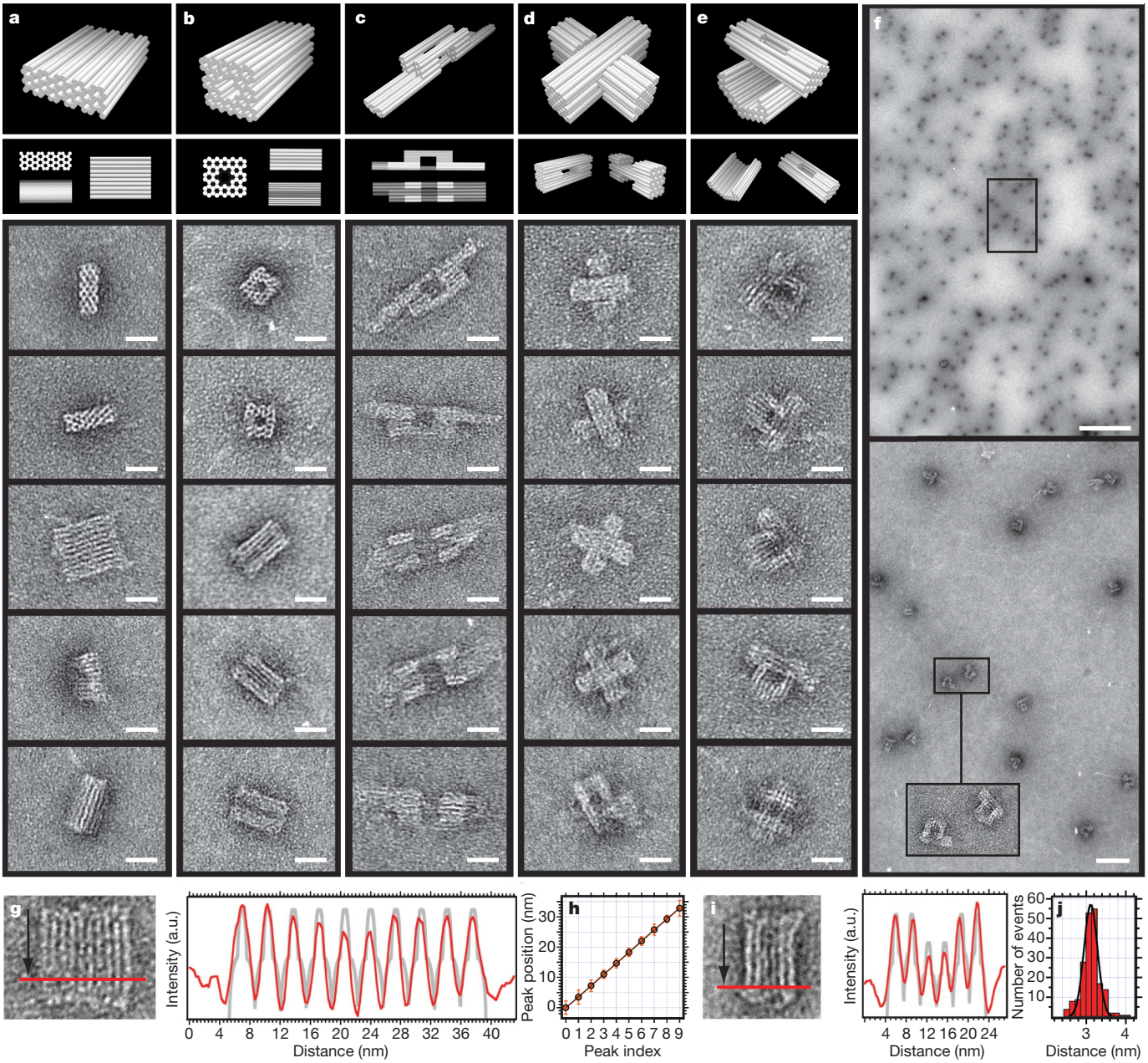


Figure 2 | Three-dimensional DNA origami shapes. The first and second rows show perspective and projection views of cylinder models, with each cylinder representing a DNA double helix. **a**, Monolith. **b**, Square nut. **c**, Railed bridge. **d**, Slotted cross. **e**, Stacked cross. Rows three to seven show transmission electron microscope (TEM) micrographs of typical particles. For imaging, samples were adsorbed (5 min) onto glow-discharged grids pretreated with 0.5 M $MgCl_2$, stained with 2% uranyl formate, 25 mM NaOH (1 min), and visualized with an FEI Tecnai T12 BioTWIN at 120 kV. **f**, Top, field of homogeneous and monodisperse stacked-cross particles. Bottom, expanded view of boxed area from above. **g**, Left, typical monolith particle. Right, integrated-intensity profile (red) of line orthogonal to the longitudinal axis of typical monolith particle, with expected profile (grey) assuming a simple homogeneous cylinder model. **h**, Left, gaussian-fitted mean peak positions (circles) in such integrated-line profiles for twenty different monolith particles as a function of peak index. The observed mean peak-to-peak distance was 3.65 nm (±0.2 nm s.d., ±0.01 nm standard error of the mean, s.e.m.). This peak-to-peak distance should correspond to 1.5 times the effective diameter d of individual double helices in the monolith structure, hence $d = 2.4$ nm. Solid line is a linear fit with a slope of 3.65 nm from peak to peak, corroborating equidistant arrangement of helices across the entire particle width. Error bars (red) indicate mean width of the peaks. Slightly higher variations in peak width at the edges of the particles are most likely due to frayed edges (compare with particles in **a** and **g**). **i**, Analysis as in **g** repeated for the square-nut shape. **j**, Histogram of gaussian-fitted peak-to-peak distances as found for the square-nut particles, with the mean value at 3.18 nm (±0.2 nm s.d., ±0.01 nm s.e.m.), indicating an effective diameter of 2.1 nm per individual double helix. a.u., arbitrary units. Scale bars: **a**–**e**, 20 nm; **f**, 1 μm (top), 100 nm (bottom).

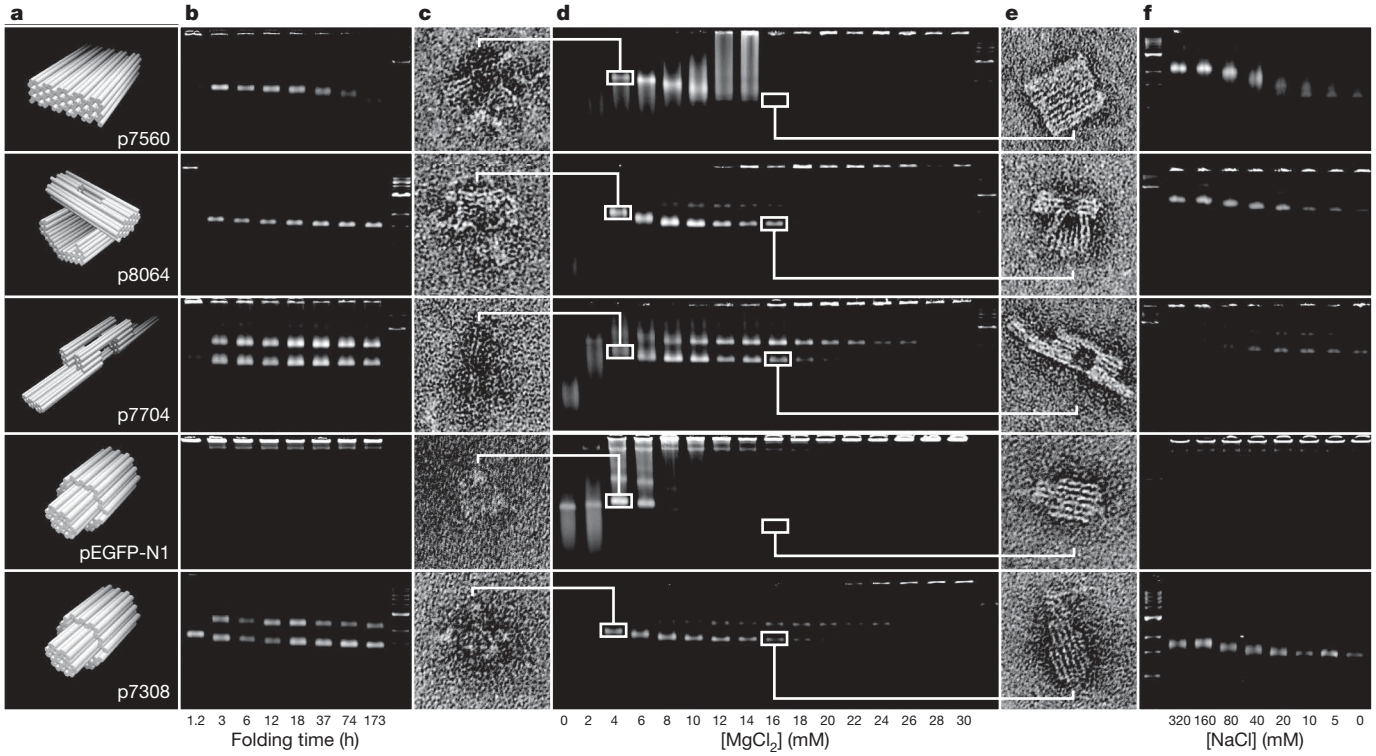


Figure 3 | Gel and TEM analysis of folding conditions for three-dimensional DNA origami. **a**, Cylinder models of shapes: monolith, stacked cross, railed bridge, and two versions of genie bottle, with corresponding scaffold sequences. Labels indicate the source of scaffold used for folding the object (for example, p7560 is an M13-based vector of length 7,560 bases). **b**, Shapes were folded in 5 mM Tris + 1 mM EDTA (pH 7.9 at 20 °C) and 16 mM $MgCl_2$ and analysed by gel electrophoresis (2% agarose, 45 mM Tris borate + 1 mM EDTA (pH 8.3 at 20 °C), 11 mM $MgCl_2$) using different thermal-annealing ramps. For the 1.2 h ramp, the temperature was lowered from 95 °C to 20 °C at a rate of 1.6 min $°C^{-1}$. For the 3 h, 6 h, 12 h, 18 h, 37 h, 74 h and 173 h ramps, the temperature was lowered from 80 °C to 60 °C at 4 min $°C^{-1}$, and then from 60 °C to 24 °C at rates of 5, 10, 20, 30, 60, 120 or 280 min $°C^{-1}$, respectively). **c**–**e**, TEM and gel analysis of influence of $MgCl_2$ concentration on folding quality. **c**, The fastest-migrating bands in the 4 mM $MgCl_2$ lanes were purified and imaged, revealing gross folding defects. **d**, Shapes were folded with a 173 h ramp in 5 mM Tris + 1 mM EDTA (pH 7.9 at 20 °C) and $MgCl_2$ concentrations varying from 0 to 30 mM. **e**, As in **c**, leading bands were purified from the 16 mM $MgCl_2$ lanes and found to exhibit higher-quality folding when analysed by TEM. **f**, Excess NaCl inhibits proper folding. Shapes were folded with 173 h ramp in 5 mM Tris + 1 mM EDTA (pH 7.9 at 20 °C), 16 mM $MgCl_2$, and varying NaCl concentrations.

with a length of 7,308 bases, where 2,800 bases of the scaffold strand were left unpaired and dangling from the neck of the bottle (reminiscent of wisps of smoke in TEM images), while the other scaffold sequence used was the 4,733-base forward strand of an expression vector encoding the enhanced green fluorescent protein (pEGFP-N1, Clontech). Folding of the same shape under identical conditions gave superior yield with the M13-based scaffold sequence. Some folding success could be achieved with the pEGFP-N1 scaffold sequence when much higher scaffold and staple concentrations were used. One striking difference between the two scaffold sequences is that the M13 base composition is 43% cytosines and guanines whereas the pEGFP-N1 base composition is 53% cytosines and guanines. Higher levels of GC base pairs might lead to a greater incidence of mispairing during folding and a slower rate of unpairing in misfolded intermediates, which could explain why folding was more difficult with the pEGFP-N1 scaffold sequence. On the other hand, local sequence diversity is potentially greatest at 50% GC content, and so a scaffold sequence with GC content that is very low might not be well-suited for DNA origami. Systematic studies will be required in the future to determine the optimal base composition.

Hierarchical assembly of DNA-origami nanostructures can be achieved by programming staple strands to bridge separate scaffold strands. Figure 4a shows the stacked cross programmed to polymerize along the long axes of the DNA helices of the pod domain. The scaffold loops on the ends of the object were programmed with a length such that they form properly spaced scaffold crossovers in the presence of bridging staple strands that link the two ends of the objects. This induces head-to-tail polymerization. Shown are filaments that adsorbed on the grid in two different orientations to illustrate the periodic presentation of the C-shaped domain perpendicular to the filament axis at a periodicity of 41 nm (±3 nm s.d. over a 33mer), corresponding to a length per base pair of 0.33 nm (±0.02 nm s.d.).

Figure 4b shows a wireframe DNA-origami nanostructure whose struts are six-helix-bundle nanotubes (strand diagrams in Supplementary Note S2). A single scaffold strand is folded into a branched tree that links two pairs of half-struts internally to produce a double triangle (Fig. 4b). This operation is repeated twice more with two completely different sets of staple strands, based on cyclic permutation of the same 8,100-base scaffold sequence through the architecture of the double-triangle monomer. This produces three chemically distinct double-triangle monomers that vary according to the sequences displayed at various positions. Every double triangle displays ten terminal branches presenting scaffold and staple sequences that are programmed to pair specifically with five terminal branches each on the two other double triangles (Fig. 4c). When the three species are mixed together, heterotrimers in the shape of a wireframe icosahedron with a diameter of about 100 nm are formed (Fig. 4d, and gel in Supplementary Note S1). The majority of particles visualized by transmission electron microscopy have missing struts, owing either to incomplete folding or to particle flattening and collapse, commonly seen for spherical or cylindrical particles prepared by negative-stain protocols[27].

Previously, scaffolded DNA origami was employed to create flat structures containing dozens of helices and nanotubes containing six helices[9,28,29]. The present work generalizes this method into three dimensions by folding helices on a honeycomb lattice. Using caDNAno[21],

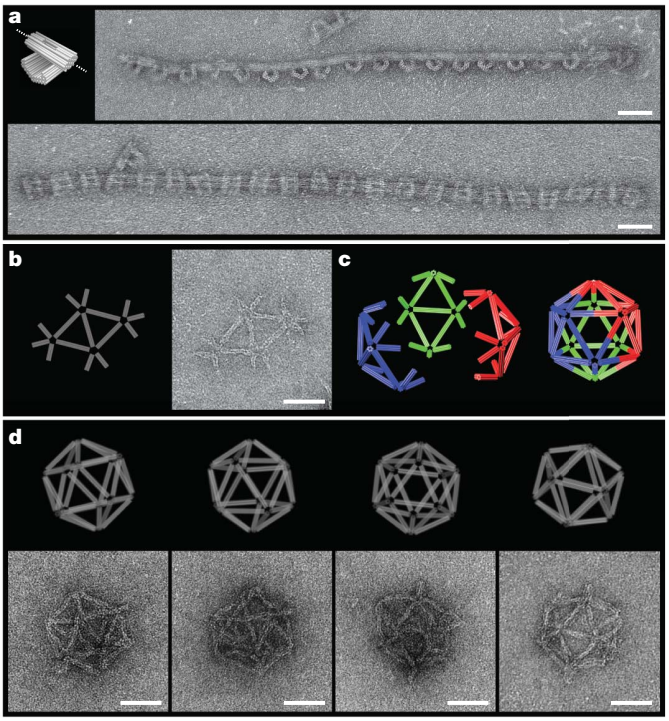


Figure 4 | Two-step hierarchical assembly of larger three-dimensional structures and polymers. a, Left panel, Cylinder model of stacked-cross monomer (Fig. 2e), with dotted line indicating direction of assembly. Right panels, typical TEM micrographs showing stacked-cross polymers. Purified stacked-cross samples were mixed with a fivefold molar excess of connector staple strands in the presence of 5 mM Tris + 1 mM EDTA (pH 7.9 at 20 °C), 16 mM $MgCl_2$ at 30 °C for 24 h. Monomers were folded in separate chambers, purified, and mixed with connector staple strands designed to bridge separate monomers. **b**, Cylinder model (left) and transmission electron micrograph (right) of a double-triangle shape comprised of 20 six-helix bundle half-struts. **c**, Heterotrimerization of the icosahedra was done with a 1:1:1 mixture of the three unpurified monomers at 50 °C for 24 h. **d**, Orthographic projection models and TEM data of four icosahedron particles. Scale bars in **a**, **b** and **d**: 100 nm.

staple sequences for folding newly conceived objects can be generated quickly. Design, acquisition of commercially synthesized staple strands, thermal folding, and initial transmission-electron-microscopic imaging can all be completed in as little as two weeks.

Improvements in the rate and yield of folding will be critical for enabling the robust assembly of larger and more complicated DNA nanostructures. Potential steps in this direction include enzymatic synthesis for higher-quality staple strands, artificial scaffold sequences that are more amenable to robust folding, folding with formamide dilution instead of thermal ramps to reduce thermal damage to the DNA[29], and hierarchical assembly with monomer architectures that have been identified as being particularly well-behaved.

Three-dimensional origami structures should expand the range of possible applications with an increased range of spatial positioning that is not accessible by flat structures, including those requiring encapsulation or space-filling functionalities. For example, many natural biosynthetic machines, such as polymerases, ribosomes, chaperones, and modular synthases, use three-dimensional scaffolding to control assembly of complex products. Similar capabilities for synthetic machines are thus more accessible with this convenient, generalizable facility to fabricate custom-shaped three-dimensional structures from DNA.

Received 16 December 2008; accepted 24 March 2009.

1. Whitesides, G. M., Mathias, J. P. & Seto, C. T. Molecular self-assembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. *Science* **254,** 1312–1319 (1991).
2. Seeman, N. C. Nucleic acid junctions and lattices. *J. Theor. Biol.* **99,** 237–247 (1982).
3. Fu, T. J. & Seeman, N. C. DNA double-crossover molecules. *Biochemistry* **32,** 3211–3220 (1993).
4. Li, X. J., Yang, X. P., Qi, J. & Seeman, N. C. Antiparallel DNA double crossover molecules as components for nanoconstruction. *J. Am. Chem. Soc.* **118,** 6131–6140 (1996).
5. Seeman, N. C. DNA in a material world. *Nature* **421,** 427–431 (2003).
6. Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394,** 539–544 (1998).
7. Yan, H., Park, S. H., Finkelstein, H., Reif, J. H. & LaBean, T. H. DNA-templated self-assembly of protein arrays and highly conductive nanowires. *Science* **301,** 1882–1884 (2003).
8. Rothemund, P. W. *et al.* Design and characterization of programmable DNA nanotubes. *J. Am. Chem. Soc.* **126,** 16344–16352 (2004).
9. Mathieu, F. *et al.* Six-helix bundles designed from DNA. *Nano Lett.* **5,** 661–665 (2005).
10. Liu, D., Park, S. H., Reif, J. H. & LaBean, T. H. DNA nanotubes self-assembled from triple-crossover tiles as templates for conductive nanowires. *Proc. Natl Acad. Sci. USA* **101,** 717–722 (2004).
11. Yin, P. *et al.* Programming DNA tube circumferences. *Science* **321,** 824–826 (2008).
12. Goodman, R. P. *et al.* Rapid chiral assembly of rigid DNA building blocks for molecular nanofabrication. *Science* **310,** 1661–1665 (2005).
13. Chen, J. H. & Seeman, N. C. Synthesis from DNA of a molecule with the connectivity of a cube. *Nature* **350,** 631–633 (1991).
14. Zhang, Y. & Seeman, N. C. The construction of a DNA truncated octahedron. *J. Am. Chem. Soc.* **116,** 1661–1669 (1994).
15. He, Y. *et al.* Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* **452,** 198–201 (2008).
16. Zhang, C. *et al.* Conformational flexibility facilitates self-assembly of complex DNA nanostructures. *Proc. Natl Acad. Sci. USA* **105,** 10665–10669 (2008).
17. Shih, W. M., Quispe, J. D. & Joyce, G. F. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature* **427,** 618–621 (2004).
18. Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295,** 2418–2421 (2002).
19. Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302 (2006).
20. Yan, H., LaBean, T. H., Feng, L. & Reif, J. H. Directed nucleation assembly of DNA tile complexes for barcode-patterned lattices. *Proc. Natl Acad. Sci. USA* **100,** 8103–8108 (2003).
21. Douglas, S. M. *et al.* Rapid prototyping of three-dimensional DNA-origami shapes with caDNAno. *Nucleic Acids Res.* (in the press).
22. Mandelkern, M., Elias, J. G., Eden, D. & Crothers, D. M. The dimensions of DNA in solution. *J. Mol. Biol.* **152,** 153–161 (1981).
23. Diekmann, S. & Lilley, D. M. J. The anomalous gel migration of a stable cruciform: temperature and salt dependence, and some comparisons with curved DNA. *Nucleic Acids Res.* **15,** 5765–5774 (1987).
24. Budker, V., Trubetskoy, V. & Wolff, J. A. Condensation of nonstoichiometric DNA/polycation complexes by divalent cations. *Biopolymers* **83,** 646–657 (2006).
25. Hibino, K. *et al.* Na+ more strongly inhibits DNA compaction by spermidin(3+) than K+. *Chem. Phys. Lett.* **426,** 405–409 (2006).
26. Garcia, H. G. *et al.* Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers* **85,** 115–130 (2007).
27. Harris, J. R., Gerber, M., Gebauer, W., Wernicke, W. & Markl, J. Negative stains containing trehalose: application to tubular and filamentous structures. *Microsc. Microanal.* **2,** 43–52 (1996).
28. Douglas, S. M., Chou, J. J. & Shih, W. M. DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proc. Natl Acad. Sci. USA* **104,** 6644–6648 (2007).
29. Jungmann, R., Liedl, T., Sobey, T. L., Shih, W. & Simmel, F. C. Isothermal assembly of DNA origami structures using denaturing agents. *J. Am. Chem. Soc.* **130,** 10062–10063 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank X. Su for assistance in cloning M13-based scaffold sequences and G. Hess for pilot studies on the railed-bridge design. This work was supported by a Claudia Adams Barr Program Investigator grant, a Wyss Institute for Biologically Inspired Engineering at Harvard grant, and an NIH New Investigator grant (1DP2OD004641-01) to W.M.S., a Humboldt Fellowship to H.D., Deutscher Akademischer Austauschdienst (DAAD) Fellowship to T.L., and Swedish Science Council (Vetenskapsrådet) Fellowship to B.H.

Author Contributions S.M.D. designed the monolith and square nut, and provided caDNAno software support; H.D. designed the stacked cross; T.L. designed the railed bridge; B.H. designed the slotted cross; F.G. designed the genie bottle; W.M.S. designed the icosahedron; S.M.D. and W.M.S. developed the honeycomb-pleated-origami design rules; H.D., S.M.D., T.L., B.H. and W.M.S. optimized the folding and imaging conditions. All authors collected and analysed data and contributed to preparing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to W.M.S. (william_shih@dfci.harvard.edu).

LETTERS

Microbial habitability of the Hadean Earth during the late heavy bombardment

Oleg Abramov[1] & Stephen J. Mojzsis[1]

Lunar rocks[1,2] and impact melts[3], lunar[4] and asteroidal meteorites[5], and an ancient martian meteorite[6] record thermal metamorphic events with ages that group around and/or do not exceed 3.9 Gyr. That such a diverse suite of solar system materials share this feature is interpreted to be the result of a post-primary-accretion cataclysmic spike in the number of impacts commonly referred to as the late heavy bombardment (LHB)[1–7]. Despite its obvious significance to the preservation of crust and the survivability of an emergent biosphere, the thermal effects of this bombardment on the young Earth remain poorly constrained. Here we report numerical models constructed to probe the degree of thermal metamorphism in the crust in the effort to recreate the effect of the LHB on the Earth as a whole; outputs were used to assess habitable volumes of crust for a possible near-surface and subsurface primordial microbial biosphere. Our analysis shows that there is no plausible situation in which the habitable zone was fully sterilized on Earth, at least since the termination of primary accretion of the planets and the postulated impact origin of the Moon. Our results explain the root location of hyperthermophilic bacteria in the phylogenetic tree for 16S small-subunit ribosomal RNA[8], and bode well for the persistence of microbial biospheres even on planetary bodies strongly reworked by impacts.

Analyses of lunar crust samples[1,2] and impact melts[3] returned by the Apollo and Luna missions, as well as from lunar meteorites[4], indicate thermal metamorphism by events that typically group around and/or do not exceed 3.9 Gyr in age. This has been interpreted to be the consequence of a pronounced increase in the number of impacts ~3.9 Gyr ago within a period of 20 to 200 Myr (refs 2, 7), in what has been termed the lunar cataclysm[2] or, more commonly, the late heavy bombardment[7]. Shock ages suggestive of the LHB have also been documented in asteroidal meteorites[5], and the only known sample of the martian crust dating from earlier than 4.0 Gyr ago, meteorite ALH 84001[6]. However, evidence of the LHB on the Earth has apparently been all but erased[9] by crustal recycling processes.

Habitats at the immediate surface of the early Earth were almost certainly repeatedly destroyed by the LHB. At the same time, however, new subsurface habitats would have formed from impact-induced hydrothermal systems[10], which perhaps provided sanctuary to existing life or may even have been the crucible of its origin[11]. The cessation of the LHB coincides very well with the earliest evidence for marine sediments[12,13] and proxy stable isotopic evidence for life existing on the Earth ~3.83 Gyr ago[14,15]. Molecular phylogenetic evidence in the form of 16S small-subunit rRNA strongly suggests that all terrestrial life arose from a common ancestral population akin to present-day thermophilic or hyperthermophilic organisms[8]. This observation has been used to implicate the bombardment epoch as a means of effectively creating an impact-generated thermal bottleneck for the biosphere[16]. It has also been postulated that the energy liberated during the LHB may have precluded the continuous survival of incipient life[17] in one or more 'impact frustrations', and disrupted the crust to such a degree that no Earth rocks survive from before ~3.8 Gyr ago[18]. However, Earth rocks that coincide with or pre-date the LHB are known, and date from as early as 4.03 Gyr ago[19]. Furthermore, geochemical evidence from terrestrial zircons older than 4.0 Gyr points to there having been liquid water, crustal recycling, granitoid crust and low-temperature plate boundary processes throughout the Hadean eon (4.38–3.85 Gyr ago)[20,21]. On the basis of this new perspective of the very early Earth, it makes sense to address whether a biosphere that may have arisen in the Hadean could have survived through the LHB.

To explore the thermal state and habitability of the early Earth during the LHB, we constructed thermal models that incorporate (1) new studies of impact cratering records of the Moon and the terrestrial planets, and the size distributions of asteroid populations[22]; (2) data from a new class of early-Solar-System dynamical models that successfully reproduce impact rates during the bombardment as defined by the lunar and meteoritic record[23]; and (3) powerful numerical methods that explore the thermal response of the lithosphere to impacts of the severity and frequency ascribed to the bombardment (Fig. 1).

Impactors that bombarded the Earth and Moon are thought to have been dominated by main-belt asteroids[22], and the size–frequency distribution of the asteroid belt is unlikely to have changed significantly since that time[24]. The total mass delivered to the Earth during the LHB has been estimated at 1.8×10^{20} kg on the basis of dynamical modelling[23] and 2.2×10^{20} kg on the basis of the lunar cratering record[25,26]. For the purposes of this work, we adopt the average value of 2.0×10^{20} kg and we use the size–frequency distribution of the asteroid belt normalized to this total mass value. The duration of the LHB in this analysis is taken to be ~100 Myr, although other values (for example 10 Myr) were also investigated.

It is likely that a relatively few very large impactors were responsible for most of the mass (and energy) delivered during the LHB (Fig. 2). Our model predicts ~90 impactors 50 km or more in diameter, which formed basins ~1,000 km or more in diameter. These impacts would have been temporally separated by over 1 Myr, on average, over the course of a 100-Myr-long bombardment, and would have resurfaced less than 25% of the Earth's surface. We evaluated the immediate thermal effects of these (and smaller) impactors on the bulk lithosphere. The fraction of the Earth's lithosphere that experienced a given temperature increase during the LHB is shown in Fig. 3. The results of this study indicate that most of the crust was not melted or thermally metamorphosed to a significant degree, with less than 10% experiencing a temperature increase of over 500 °C. The total mass delivered would have to be approximately tenfold greater than that in our baseline model for most of the crustal volume to have undergone melting. Lithospheric thickness has no significant influence on these results.

[1]University of Colorado, Department of Geological Sciences, 2200 Colorado Avenue, UCB 399, Boulder, Colorado 80309-0399, USA.

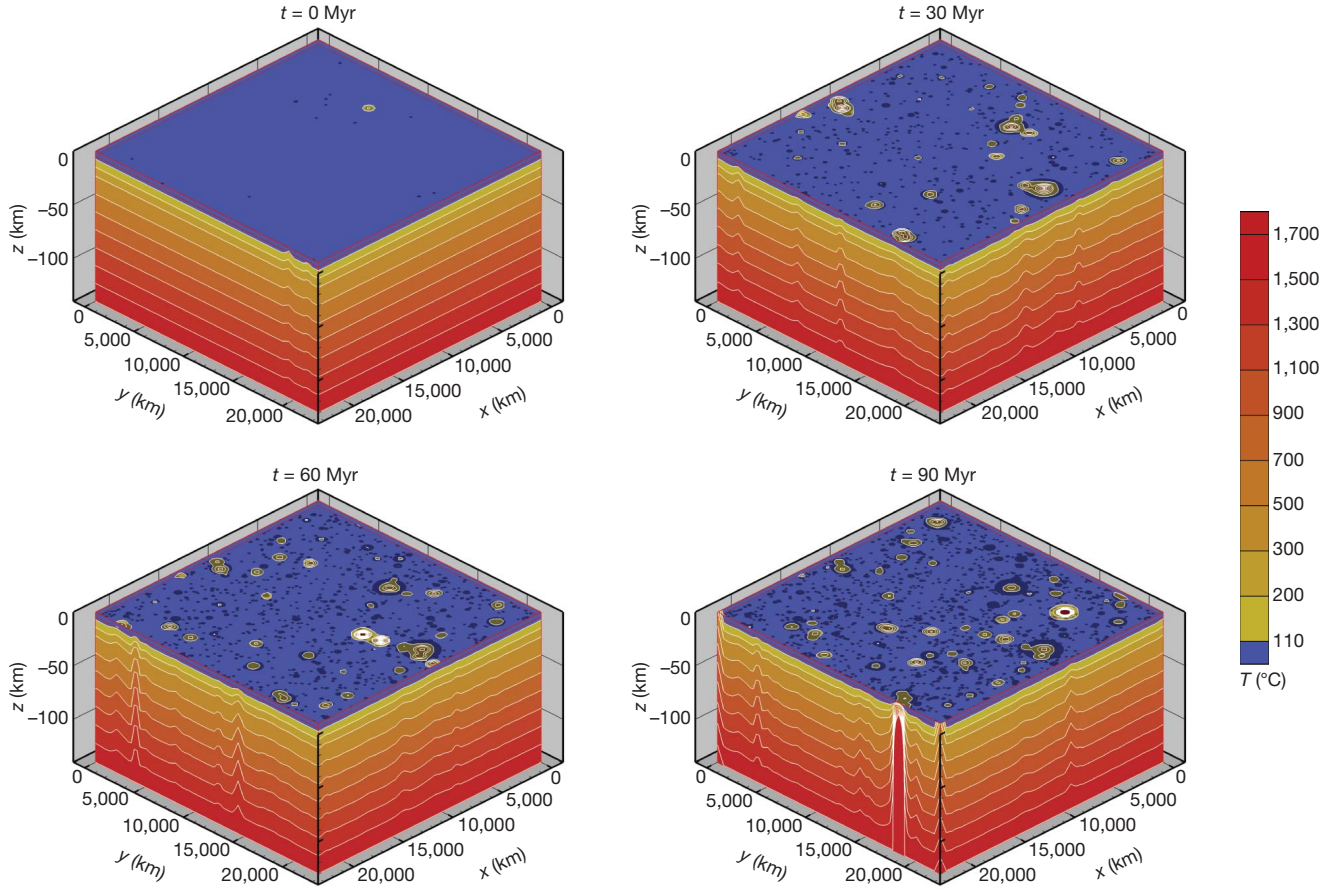


Figure 1 | A three-dimensional thermal model representing the Earth's lithosphere at various times during the LHB in our baseline scenario. Only impactors larger than 10 km in diameter are included in this plot. The upper surface shows temperatures at a depth of 4 km. Dark areas indicate crater imprints.

Whereas very large basin-forming impacts deposit most of their energy in the deep lithosphere and the mantle, smaller impacts deposit most of their energy in the potentially habitable top 4 km of the crust. Thus, for a more complete habitability evaluation, we performed an assessment of the near-surface thermal effects of a wide range of impactors. The results of this analysis show that smaller impactors (diameter, 1–10 km) were as important as much larger basin formers (≥100 km) in terms of sterilization (heating above 110 °C) of the near-surface crust (Table 1). Nonetheless, large craters are more biologically significant because their near-surface crust takes orders of magnitude longer to cool (~10^7 yr for a 2,000-km impact basin versus ~10^5 yr for a 20-km crater), which results in long-lived hydrothermal systems[27]. The outcomes presented in Table 1 also suggest that even if all LHB impacts had occurred simultaneously, Earth still would not have been sterilized.

Habitable volumes for mesophile (20–50 °C), thermophile (50–80 °C) and hyperthermophile (80–110 °C) microbial populations in the near-surface crust (the upper ~4 km) for a 100-Myr-long

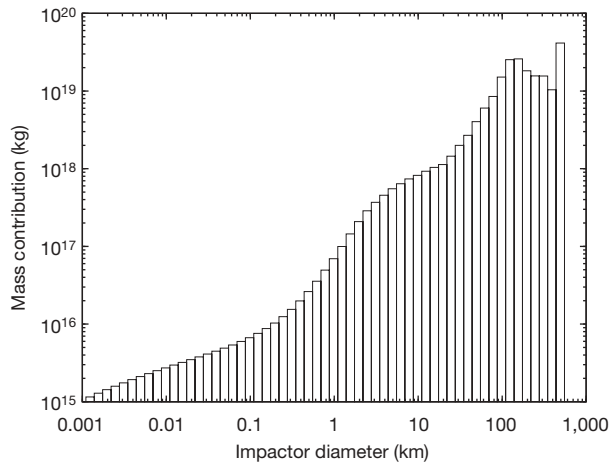


Figure 2 | Impactor mass distribution for the Earth during the LHB. Distribution derived from the main-belt size–frequency distribution[24] and estimates of the total mass delivered[23,25,26].

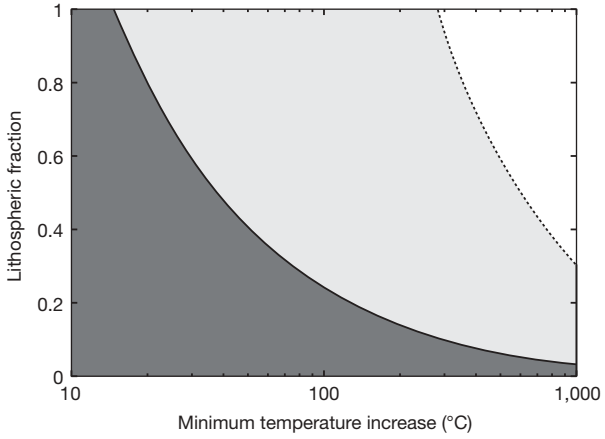


Figure 3 | Immediate thermal effects of impacts on the lithosphere. Fraction of the Earth's lithosphere to experience a given temperature increase as a result of the LHB. The solid line represents the baseline model and the dashed line represents the model with tenfold delivered mass. Lithospheric thickness has no significant effect on these results.

Table 1 | Immediate thermal effects of impacts on the habitable zone

Impactor diameter range (km)	Number of impacts	Percentage of habitable zone sterilized
≥100	33	13%
10–100	1,500	10%
1–10	170,000	13%
0.1–1	12,000,000	1%

Percentage of habitable zone (~4 km below the surface) exposed to temperatures above 110 °C (the upper limit for hyperthermophiles) in the baseline model.

LHB are shown in Fig. 4a. We find that over the course of the LHB, there would have been a significant increase in the habitable volume for thermophiles and hyperthermophiles and a decrease in the habitable volume for mesophiles. The overall habitable volume remains approximately the same, however, because of the relatively rapid cooling of the near-surface crust even in the case of large impact basins. Habitable conditions in the near-surface volume of the crust are re-established in at most $\sim 10^5$ yr after impact.

Although the LHB in our main model is insufficient to extinguish microbial life in the Hadean Earth's habitable volume of crust, the model does not explicitly incorporate thermal shock from a global layer of hot ejecta and rock-vapour rainout following a basin-forming impact[28]. However, the maximum sterilization depth for such a process is only a few hundred metres[28] and is further reduced or eliminated by vaporization of groundwater, hydrothermal circulation, the raining out of impact-vaporized and hydrothermally vented water, and the presence of oceans (Supplementary Information). Finally, the largest impactor in our baseline model (~300 km in diameter) is insufficient to vaporize the oceans[28]. As the duration of the LHB is still uncertain, one of our model test cases was an LHB with a duration of 10 Myr, with the total delivered mass held at 2.0×10^{20} kg. The output shows a pronounced increase in the hyperthermophile habitable volume and a decrease in the mesophile habitable volume (Fig. 4b). The mesophile time–temperature curve approaches the thermophile curve and crosses it at several points.

Increasing the total mass delivered by a factor of ten (to 2.0×10^{21} kg), with a 100-Myr-long bombardment, results in a marked enhancement of the hyperthermophile habitable volume and a sharp decrease in the mesophile habitable volume (Fig. 4c). The calculated total habitable volume of the Hadean crust (1.7×10^9 km^3) at the end of the LHB is slightly lower than the habitable volume (2.1×10^9 km^3) at the start of the simulation. A

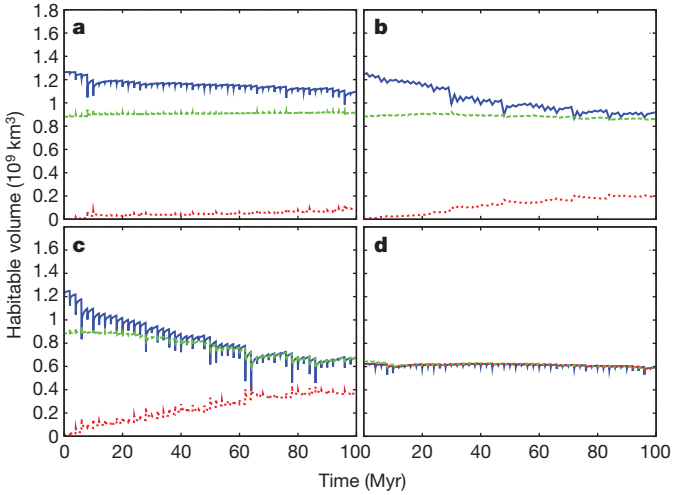


Figure 4 | Global habitable volumes. **a**, Habitable volumes for mesophiles (20–50 °C, blue), thermophiles (50–80 °C, green) and hyperthermophiles (80–110 °C, red) during a 100-Myr-long LHB. **b**, Habitable volumes during a 10-Myr-long LHB with the same delivered mass. **c**, Habitable volumes during a 100-Myr-long LHB with tenfold delivered mass. **d**, Habitable volumes during a 100-Myr-long LHB with the geothermal temperature gradient doubled.

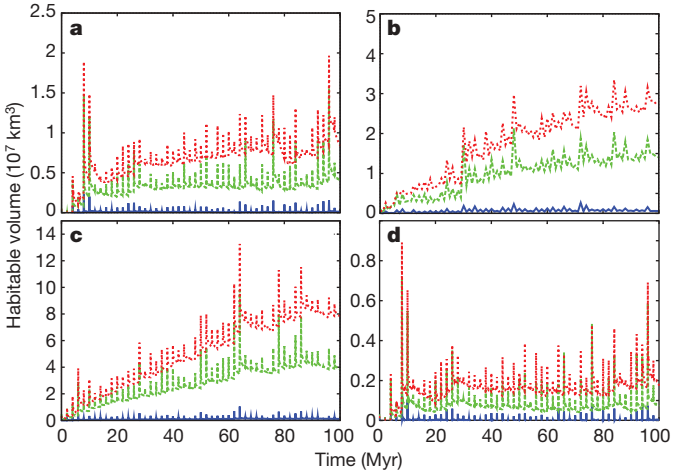


Figure 5 | Global habitable volumes in hydrothermal environments. **a**, Habitable volumes for mesophiles (20–50 °C, blue), thermophiles (50–80 °C, green) and hyperthermophiles (80–110 °C, red) during a 100-Myr-long LHB in active hydrothermal environments. **b**, Hydrothermal habitable volumes during a 10-Myr-long LHB with the same delivered mass. **c**, Hydrothermal habitable volumes during a 100-Myr-long LHB with tenfold delivered mass. **d**, Hydrothermal habitable volumes during a 100-Myr-long LHB with the geothermal temperature gradient doubled.

similar plot of habitable volume can be produced by doubling the impact speed from 20 to 40 km s^{-1} (Supplementary Information).

Doubling the geothermal temperature gradient from 12 °C km^{-1} to 24 °C km^{-1} results in approximately equal habitable volumes (~6×10^8 km^3) for mesophiles, thermophiles and hyperthermophiles throughout the LHB (Fig. 4d). This is due to their initial equal volumes in the habitable zone and the relatively rapid linearization of the vertical temperature gradients after each impact. We also investigated an extreme endmember case in which the surface temperature is 50 °C, the geothermal gradient is 48 °C km^{-1} and the total delivered mass is 2.0×10^{22} kg (100 times the baseline), and found that this is insufficient to extinguish subsurface microbial life owing to the relatively rapid cooling of the near-surface crust following an impact (Supplementary Information).

If the subsurface is saturated with water and hydrothermal circulation is established, heat is lost from the upper boundary up to ten times faster, and habitable conditions are re-established up to an order of magnitude more rapidly after crater formation[29]. This is particularly relevant if the impact occurs in the ocean. For craters ~200 km in diameter, colonization by thermophiles in the central regions is possible after ~20,000 yr (ref. 29). This observation provides further support to the conclusion that subsurface microbial life could have persisted throughout the bombardment.

We also estimated habitable volumes in impact-induced temperature anomalies that are likely to drive hydrothermal activity (Fig. 5). For the purposes of this study, a hydrothermal environment is defined as a thermally anomalous environment with a geothermal temperature gradient in excess of double the normal value. Comparison of Fig. 4 and Fig. 5 shows that habitable volumes in active hydrothermal environments are a relatively small fraction of the total—in no case is the habitable volume dominated by hydrothermal environments. This would seem to argue against the idea of an impact bottleneck in which only thermophiles survived[28], and instead supports the hypothesis that widespread hydrothermal activity during the LHB was conducive to life's emergence and early diversification[30].

METHODS SUMMARY

Our stochastic cratering model was used to populate all or part of the Earth's surface with craters within a probability field of constraints established from the models and observations cited above. The thermal field of each crater was introduced into a three-dimensional block model of the Earth's lithosphere and

allowed to cool by conduction in the subsurface and radiation/convection at the atmosphere interface (Fig. 1). We used the computer code HEATING, which is a general-purpose, three-dimensional, finite-difference heat-transfer program written and maintained at Oak Ridge National Laboratory, USA, that solves transient heat-conduction problems and contains a library of thermal properties of geologically relevant materials, such as basalt and granite. The physical and thermal properties of each target rock (for example thermal conductivity, density and specific heat) are temperature dependent. Target crust was allowed by the code to undergo changes of phase, permitting the inclusion of impact melt. Boundary conditions included a bottom boundary with a prescribed heat flux and a surface–atmosphere interface where heat is transferred out of the system by convection and radiation. The lateral boundaries of the block model are defined as continuous 'wrap arounds', so that all deposited heat stays in the system until it is lost through the upper boundary.

Received 12 December 2008; accepted 24 March 2009.

1. Turner, G., Cadogan, P. H. & Yonge, C. J. Argon selenochronology. *Proc. Lunar Sci. Conf.* **4,** 1889–1914 (1973).
2. Tera, F., Papanastassiou, D. A. & Wasserburg, G. J. Isotopic evidence for a terminal lunar cataclysm. *Earth Planet. Sci. Lett.* **22,** 1–21 (1974).
3. Dalrymple, G. B. & Ryder, G. $^{40}Ar/^{39}Ar$ age spectra of Apollo 15 impact melt rocks by laser step-heating and their bearing on the history of lunar basin formation. *J. Geophys. Res.* **98,** 13085–13095 (1993).
4. Cohen, B. A., Swindle, T. D. & Kring, D. A. Support for the lunar cataclysm hypothesis from lunar meteorite impact melt ages. *Science* **290,** 1754–1756 (2000).
5. Kring, D. A. & Cohen, B. A. Cataclysmic bombardment throughout the inner solar system 3.9–4.0 Ga. *J. Geophys. Res.* **107,** doi:10.1029/2001JE001529 (2002).
6. Ash, R. D., Knott, S. F. & Turner, G. A 4-Gyr shock age for a Martian meteorite and implications for the cratering history of Mars. *Nature* **380,** 57–59 (1996).
7. Ryder, G. Lunar samples, lunar accretion, and the early bombardment history of the Moon. *Eos* **71,** 313–323 (1990).
8. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
9. Trail, D., Mojzsis, S. J. & Harrison, T. M. Thermal events documented in Hadean zircons by ion microprobe depth profiles. *Geochim. Cosmochim. Acta* **71,** 4044–4065 (2007).
10. Zahnle, K. J. & Sleep, N. H. in *Comets and the Origin and Evolution of Life* (eds Thomas, P., Chyba, C. & McKay, C.) 175–208 (Springer, 1997).
11. Baross, J. A. & Hoffman, S. E. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig. Life Evol. Biosph.* **15,** 327–345 (1985).
12. Manning, C. E., Mojzsis, S. J. & Harrison, T. M. Geology, age and origin of supracrustal rocks at Akilia, West Greenland. *Am. J. Sci.* **306,** 303–366 (2006).
13. Dauphas, N. *et al.* Clues from Fe isotope variations on the origin of early Archean BIFs from Greenland. *Science* **206,** 2077–2080 (2004).
14. Mojzsis, S. J. *et al.* Evidence for life on Earth by 3,800 million years ago. *Nature* **384,** 55–59 (1996).
15. McKeegan, K. D., Kudryavtsev, A. B. & Schopf, J. W. Raman and ion microscopic imagery of graphitic inclusions in apatite from older than 3830 Ma Akilia supracrustal rocks, west Greenland. *Geology* **35,** 591–594 (2007).
16. Gogarten-Boekels, M., Hilario, E. & Gogarten, J. P. The effects of heavy meteorite bombardment on the early evolution - the emergence of the three domains of life. *Orig. Life Evol. Biosph.* **25,** 251–264 (1995).
17. Maher, K. A. & Stevenson, D. J. Impact frustration of the origin of life. *Nature* **331,** 612–614 (1988).
18. Hamilton, W. B. in *Precambrian – Conterminous United States* (eds Reed, J. C. Jr *et al.*) 597–614, 630–636 (Geol. N. Am. Vol. C-2, Geological Society of America, 1993).
19. Bowring, S. A. & Williams, I. S. Priscoan (4.00–4.03 Ga) orthogneisses from northwestern Canada. *Contrib. Mineral. Petrol.* **134,** 3–16 (1999).
20. Mojzsis, S. J., Harrison, T. M. & Pidgeon, R. T. Oxygen isotope evidence from ancient zircons for liquid water at Earth's surface 4,300 Myr ago. *Nature* **409,** 178–181 (2001).
21. Hopkins, M., Harrison, T. M. & Manning, C. E. Low heat flow inferred from >4 Gyr zircons suggests Hadean plate boundary interactions. *Nature* **456,** 493–496 (2008).
22. Strom, R. G., Malhotra, R., Ito, T., Yoshida, F. & Kring, D. A. The origin of planetary impactors in the inner solar system. *Science* **309,** 1847–1850 (2005).
23. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435,** 466–469 (2005).
24. Bottke, W. F. *et al.* The fossilized size distribution of the main asteroid belt. *Icarus* **175,** 111–140 (2005).
25. Hartmann, W. K., Ryder, G., Dones, L. & Grinspoon, D. in *Origin of the Earth and Moon* (eds Canup, R. & Righter, K.) 493–512 (Univ. Arizona Press, 2000).
26. Ryder, G., Koeberl, C. & Mojzsis, S. J. in *Origin of the Earth and Moon* (eds Canup, R. & Righter, K.) 475–492 (Univ. Arizona Press, 2000).
27. Abramov, O. & Kring, D. A. Impact-induced hydrothermal activity on early Mars. *J. Geophys. Res.* **110,** doi:10.1029/2005JE002453 (2005).
28. Sleep, N. H. & Zahnle, K. Refugia from asteroid impacts on early Mars and the early Earth. *J. Geophys. Res.* **103,** 28529–28544 (1998).
29. Abramov, O. & Kring, D. A. Numerical modeling of an impact-induced hydrothermal system at the Sudbury crater. *J. Geophys. Res.* **109,** doi:10.1029/2003JE002213 (2004).
30. Ryder, G. Mass flux in the ancient Earth-Moon system and benign implications for the origin of life on Earth. *J. Geophys. Res.* **107,** doi:10.1029/2001JE001583 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work is funded by the NASA Astrobiology Institute (through the NASA Postdoctoral Program) and the NASA Exobiology program. Office space and computer resources provided by the Department of Space Studies of the Southwest Research Institute in the early stages of this project are greatly appreciated. Reviews by E. Asphaug, as well as comments by D. Trail and T. M. Harrison, are gratefully acknowledged.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to O.A. (oleg.abramov@colorado.edu).

LETTERS

PEP1 regulates perennial flowering in *Arabis alpina*

Renhou Wang[1], Sara Farrona[1], Coral Vincent[1], Anika Joecker[1], Heiko Schoof[1], Franziska Turck[1], Carlos Alonso-Blanco[2], George Coupland[1] & Maria C. Albani[1]

Annual plants complete their life cycle in one year and initiate flowering only once, whereas perennials live for many years and flower repeatedly. How perennials undergo repeated cycles of vegetative growth and flowering that are synchronized to the changing seasons has not been extensively studied[1]. Flowering is best understood in annual *Arabidopsis thaliana*[2,3], but many closely related species, such as *Arabis alpina*[4,5], are perennials. We identified the *A. alpina* mutant *perpetual flowering 1* (*pep1*), and showed that *PEP1* contributes to three perennial traits. It limits the duration of flowering, facilitating a return to vegetative development, prevents some branches from undergoing the floral transition allowing polycarpic growth habit, and confers a flowering response to winter temperatures that restricts flowering to spring. Here we show that *PEP1* is the orthologue of the *A. thaliana* gene *FLOWERING LOCUS C* (*FLC*). The FLC transcription factor inhibits flowering until *A. thaliana* is exposed to winter temperatures[6,7], which trigger chromatin modifications that stably repress *FLC* transcription[8–11]. In contrast, *PEP1* is only transiently repressed by low temperatures, causing repeated seasonal cycles of repression and activation of *PEP1* transcription that allow it to carry out functions characteristic of the cyclical life history of perennials. The patterns of chromatin modifications at *FLC* and *PEP1* differ correlating with their distinct expression patterns. Thus we describe a critical mechanism by which flowering regulation differs between related perennial and annual species, and propose that differences in chromatin regulation contribute to this variation.

Perennial plants repeatedly cycle between vegetative and reproductive development. In temperate climates these cycles are synchronized to the changing seasons, for example, by restricting flowering to spring and summer[12] or by arresting growth in the autumn[13]. Annual and perennial plants also show differences in the behaviour of shoot meristems—groups of undifferentiated cells from which all aerial tissues are derived. In annual plants all shoot meristems initiate reproductive development at similar times, a behaviour called monocarpy. In contrast, perennials are polycarpic and maintain vegetative growth after flowering, which allows them to flower and set seed many times during their lifetime. Vegetative growth is maintained either by conserving some meristems in the vegetative state after flower initiation[14] or by reverting back to vegetative development after flowering[15].

Control of the floral transition is best understood in the annual monocarpic model species *Arabidopsis thaliana*[2,3], a member of the *Brassicaceae*. Many other *Brassicaceae* species are polycarpic perennials, and therefore the molecular mechanisms underlying the difference between monocarpic and polycarpic plants can be approached by comparing *A. thaliana* with its close relatives. Similar comparative approaches were recently used to study the development of compound leaves[16] and to analyse the basis of heavy metal tolerance[17]. The arctic-alpine perennial *Arabis alpina* is a member of the *Brassicaceae*[4,5] that has favourable characteristics for use as a model perennial species, such as being diploid and self-fertile, with a relatively small genome, and being susceptible to transformation by *Agrobacterium tumefaciens*. Here we describe a molecular-genetic analysis of the perennial traits seasonal flowering and polycarpy in *A. alpina*. We demonstrate that the *A. alpina* gene *PERPETUAL FLOWERING 1* (*PEP1*) encodes a MADS-box transcription factor that mechanistically links polycarpy and seasonal flowering. *PEP1* is the orthologue of the *A. thaliana* gene *FLOWERING LOCUS C* (*FLC*) and it is differently regulated in *A. alpina*, allowing the repeated response to seasonal cycles observed in perennial plants. This distinct pattern of regulation is related to species-specific differences in histone modifications at *PEP1*. These data provide insight into the mechanisms underlying evolution of life history in plants.

To facilitate a genetic study of flowering in *A. alpina* we first characterized the perennial growth habit and responses to environmental signals. The accession Pajares did not flower when grown continuously under long days (16 h light) or exposed to short days (8 h light) (Fig. 1a, Supplementary Fig. 1a and Methods). However, these plants flowered when exposed to low temperatures for several weeks (Fig. 1b), a treatment called vernalization. Although the flowering response to vernalization was independent of day length (Supplementary Fig. 1a), it was influenced by the duration of exposure to cold. Twelve weeks vernalization was sufficient to saturate the flowering response, causing all plants to produce fully developed inflorescences, whereas vernalization for shorter periods resulted in inflorescences that seemed to revert to vegetative growth (Fig. 1b, e and Supplementary Fig. 1b). Therefore, *Arabis alpina* Pajares plants only flower if exposed to prolonged vernalization treatments, and this response is independent of day length.

To understand the polycarpic behaviour of *A. alpina*, the fates of apical meristems of the main shoot and axillary shoots (branches) were followed before and after vernalization. The apices of the main shoot and axillary shoots formed before cold treatment produced flower buds during vernalization (Fig. 1c and Supplementary Figs 2 and 3a–d). Mature flowers emerged when plants were subsequently moved back to normal growth temperatures. Vegetative shoots developed from meristems that were not present or were still at an early developmental stage at the onset of vernalization, contributing to the polycarpic growth habit of the plant (Supplementary Fig. 3e, f). These shoots continued growing vegetatively until plants were again exposed to vernalization, inducing another round of flowering (Fig. 1d–h). Therefore, perenniality in this species involves maintenance of vegetative development after flowering, and the requirement for vernalization to induce flowering during each annual cycle.

To study the molecular mechanisms controlling flowering of *A. alpina*, mutants showing an impaired vernalization response were identified. From a total of 25,000 M_2 seedlings, two mutants were isolated that lacked obligate vernalization requirement. The most extreme early flowering mutant was called *perpetual flowering 1* (*pep1*). This mutant flowered with approximately 25 leaves under long

[1]Max Planck Institute for Plant Breeding Research, Carl von Linne Weg 10, D-50829 Cologne, Germany. [2]Departamento de Genética Molecular de Plantas, Centro Nacional de Biotecnología (Consejo Superior de Investigaciones Científicas), Cantoblanco, 28049 Madrid, Spain.

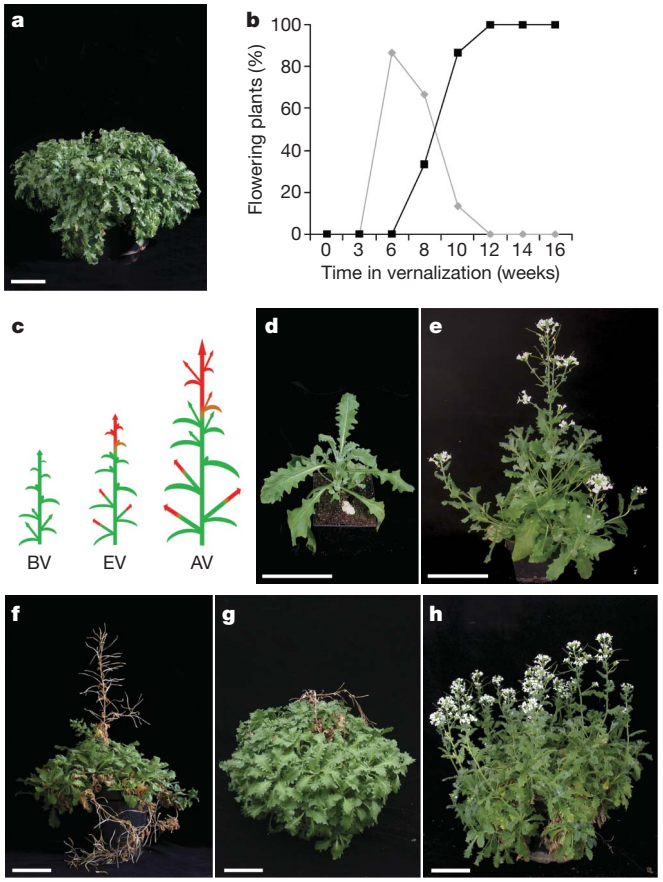


Figure 1 | *Arabis alpina* is a polycarpic perennial that requires vernalization each year to flower. **a**, An *A. alpina* (accession Pajares) plant grown for 3 and a half years without vernalization. The plant has not flowered. **b**, Quantification of the duration of vernalization required for flowering of the Pajares accession. Plants were grown for 8 weeks in long days, vernalized for different times and subsequently grown in long days. Shown in grey are plants that responded partially to vernalization by producing a few flowers but also producing vegetative characteristics on the inflorescence (see also Supplementary Fig. 1b). Shown in black are plants that completely responded to vernalization and produced fully developed inflorescences. Twelve plants were scored at each time point. **c**, Diagram showing differential response of shoots to vernalization in *A. alpina*. The main shoot and axillary shoots are vegetative (green) before vernalization and become reproductive (red) by the end of the vernalization treatment. Axillary shoots that arise during or after vernalization remain vegetative (green). BV, before vernalization, plants grown for 8 weeks in long days; EV, end of a 12-week vernalization treatment under short days at 4 °C; AV, after vernalization, plants grown under long days. **d–h**, *Arabis alpina* does not die after flowering. The reproductive part of each shoot sets seed and senesces whereas vegetative shoots maintain vegetative growth indefinitely until exposed to another round of vernalization. Pictures from the same *A. alpina* plant were taken at successive stages of its life cycle: after 8 weeks in long days (**d**), vernalized for 12 weeks and then transferred to long days for 3 weeks (**e**), after 14 weeks in long days (**f**), after 25 weeks in long days (**g**), and after a new round of 12 weeks vernalization and then again for 4 weeks in long days (**h**). Scale bars, 10 cm.

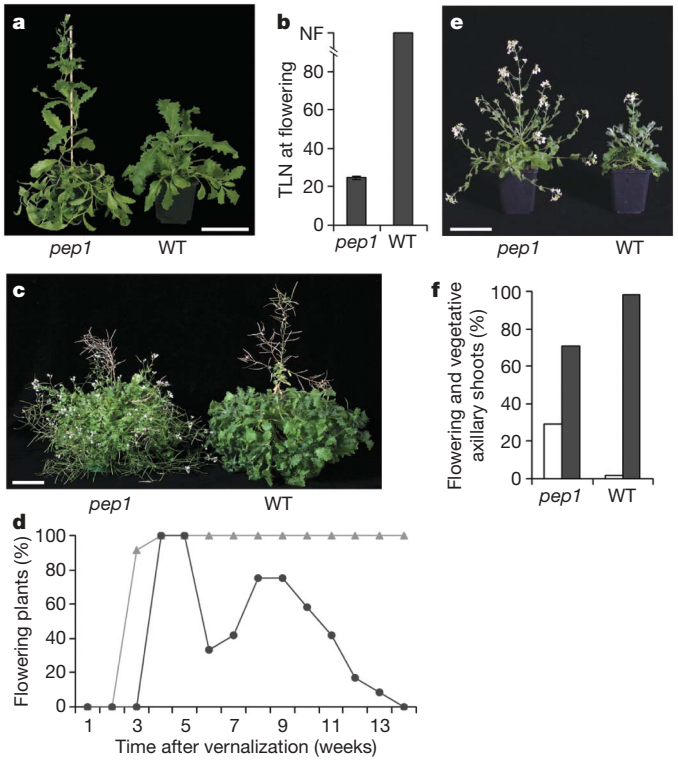


Figure 2 | *PEP1* restricts the flowering phase and enhances polycarpy of *A. alpina*. **a**, The *pep1* mutant flowers without vernalization. **b**, Flowering time of *pep1* plants compared to wild type (WT) without vernalization. NF, no flowering; TLN, total leaf number. **c**, *pep1* mutants show longer periods of flowering than wild-type plants. Plants were grown for 14 weeks under long days after 12 weeks vernalization. **d**, Duration of the flowering phase of *pep1* (grey) and wild-type plants (dark grey) after vernalization. **e**, More axillary branches undergo flowering in *pep1* plants than in wild-type as shown by plants grown for 3 weeks under long days after 12 weeks vernalization. **f**, Percentage of vegetative and flowering axillary shoots produced by *pep1* and wild-type plants. Plants were scored when the first open flower was present. Flowering shoots (white) and vegetative shoots (dark grey) are shown. In **c–f**, plants were grown under long days for 5 weeks before being vernalized for 12 weeks. $n = 12$ (**b**, **d**, **f**); scale bars, 10 cm.

days without exposure to vernalization, whereas wild-type plants never flowered under these conditions (Fig. 2a, b). In addition, the *pep1* mutant flowered continuously for at least 12 months (Supplementary Fig. 4a). To compare the duration of the flowering season of the *pep1* mutant and wild-type plants, both genotypes were vernalized and the duration of flowering was measured. Reproductive shoots in wild-type plants completed flowering 14 weeks after return to warm temperatures under long days, whereas under the same conditions *pep1* plants flowered for a further 5 months until the experiment was terminated (Fig. 2c, d). Flowering of wild-type plants was restricted to apical meristems of shoots that were present before vernalization, whereas the *pep1* mutant continued flowering from secondary axillary shoots (Supplementary Fig. 4b). Moreover, the *pep1* mutant showed altered polycarpic behaviour. More shoots were induced to flower in the *pep1* mutant because axillary shoots that emerged during or after vernalization flowered in the *pep1* mutant but not in the wild-type plants (Fig. 2e, f and Supplementary Fig. 4c, d). Our data demonstrate that *PEP1* acts at different stages in the perennial life cycle. PEP1 prevents flowering before vernalization, whereas after vernalization it acts to restrict the duration of flowering and contributes to polycarpy by blocking flowering of axillary shoots. These three functions all involve the repression of the floral transition and are likely to be mechanistically related.

The reduced requirement for vernalization in the *pep1* mutant suggested that *PEP1* may be an orthologue of an *A. thaliana* gene conferring a vernalization requirement. To test this idea we isolated the *A. alpina* orthologue of *FLOWERING LOCUS C* (*FLC*), which encodes a MADS-box transcription factor that has an important role in establishing vernalization requirement in *A. thaliana*[6,7] and probably other annual *Brassicaceae* species[18,19]. *AaFLC* is present in a genomic region showing microsynteny with the *A. thaliana* chromosomal region containing *FLC* (Fig. 3a and Supplementary Table 1), and encodes a protein more closely related to FLC than to any other *A. thaliana* protein (Supplementary Fig. 5a). The function of AaFLC was tested by expressing the *AaFLC* complementary DNA from the *CaMV35S* promoter in the accession Columbia-0, which expresses the endogenous *FLC* gene at low levels[6,7]. The transgenic plants were late

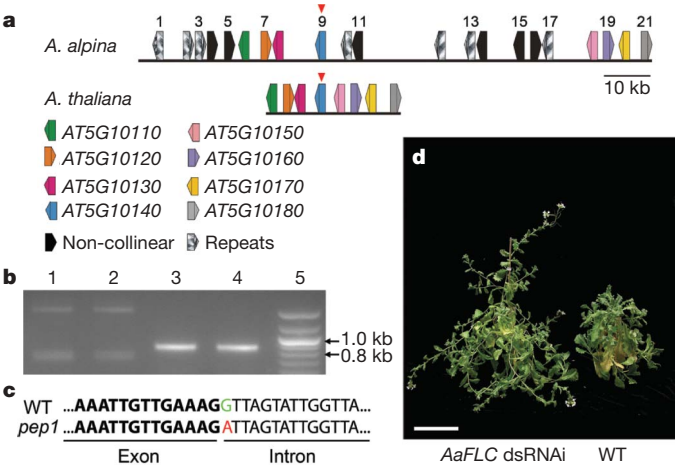


Figure 3 | *PEP1* is the *Arabis alpina* orthologue of *A. thaliana FLC*. **a**, Conserved synteny between the *Arabis alpina* (top panel) and *A. thaliana* (bottom panel) chromosomal regions containing *AaFLC* and *Arabidopsis FLC*, respectively. Vertical arrowheads show *AaFLC* and *Arabidopsis FLC* in each panel. Horizontal arrowheads illustrate genes and their orientation. Coloured arrowheads show syntenic genes, whereas black and striped arrowheads show the non-collinear genes and repeats, respectively. Genes (1–21) on the BAC containing the *AaFLC* gene are annotated in Supplementary Table 1. GenBank accession number FJ543377. **b**, Accumulation of the full-length *AaFLC* transcript is impaired in *pep1* mutant plants. Amplified products from cDNA derived from *pep1* apices (lane 1), *pep1* leaves (lane 2), wild-type apices (lane 3) and wild-type leaves (lane 4). Lane 5 contains a 100-base-pair (bp) standard (NEB). Primers were designed to amplify by PCR the full-length coding sequence of *AaFLC* (819 bp). *AaFLC* cDNA in the *pep1* mutant and wild-type plant were sequenced and the detected splice variants are shown in Supplementary Fig. 6. In the *pep1* mutant, splice variants that were larger and smaller than the fully spliced *AaFLC* ORF were detected, accounting for the differently sized cDNAs detected in lanes 1 and 2. **c**, Mutation in the *AaFLC* gene in the *pep1* mutant. The G-to-A mutation at the splice donor site of exon 3 is indicated in green (wild type, WT) and red (*pep1*). **d**, Transgenic *35S:AaFLC* double-stranded (ds)RNAi *A. alpina* Pajares plants flower without vernalization. See also Supplementary Figs 5 and 8. Scale bar, 10 cm.

flowering (Supplementary Fig. 5b, c), demonstrating that AaFLC can act as a floral repressor in *A. thaliana* as previously shown for FLC[6,7].

To test the relationship between *AaFLC* and *PEP1*, the expression of *AaFLC* messenger RNA was compared in *pep1* and wild-type plants. Full-length *AaFLC* mRNA was below the level of detection in the *pep1* mutant (Fig. 3b and Supplementary Fig. 6), and by sequencing the *AaFLC* genomic region a mutation (G-to-A) was identified at the 5′ splice junction of the third intron (Fig. 3c). In the *pep1* mutant *AaFLC* mRNAs of different sizes were detected (Fig. 3b), and therefore *AaFLC* cDNAs made from *pep1* RNA were sequenced. None of the 14 cDNAs tested contained the full *AaFLC* open reading frame (ORF; Supplementary Fig. 6). In contrast, the cDNA containing the full ORF was by far the most abundant form detected in wild-type *A. alpina* plants (Fig. 3b and Supplementary Fig. 6). These data support the idea that the mutation present in *AaFLC* in the *pep1* mutant impairs production of the properly spliced mRNA.

We further tested whether reduced *AaFLC* activity was the cause of the *pep1* mutant phenotype. First, genetic linkage between *pep1* and the mutation in *AaFLC* was tested in an F_2 population constructed by backcrossing the *pep1* mutant to wild type. Under continuous long days, the ratio of flowering to non-flowering F_2 plants was approximately 1:3 (Supplementary Fig. 7). Genotypic data demonstrated that all of the flowering plants were homozygous for the mutation in *AaFLC*, and that all non-flowering plants were either heterozygous or homozygous for the wild-type allele (Supplementary Fig. 7). This experiment demonstrated that *pep1* is genetically linked to *AaFLC*. Second, transgenic plants, in which the level of *AaFLC* mRNA was reduced by RNA interference (RNAi), were generated. These plants showed the characteristic phenotypes of *pep1* mutant plants. They flowered without vernalization, showed a longer duration of flowering and produced more flowering branches compared to the wild type (Fig. 3d and Supplementary Figs 5d, e and 8). These experiments further indicated that the *pep1* mutant phenotype is caused by impaired *AaFLC* activity and therefore *AaFLC* will be referred to as *PEP1*.

In *A. thaliana*, *FLC* transcription is repressed during vernalization, and after return to normal growth temperatures *FLC* repression is maintained[6–8,10]. The behaviour of *pep1* mutants suggested that *PEP1* functions before and after vernalization. To determine the *PEP1* expression pattern in *A. alpina*, apices and leaves of wild-type plants were analysed before, during and after vernalization using quantitative PCR with reverse transcription (qRT–PCR). During vernalization, *PEP1* transcript levels decreased markedly in shoot apices and leaves, and reached much lower levels at the end of the low temperature treatment, as observed in *A. thaliana*. However, after return to normal growth temperatures, *PEP1* mRNA levels increased in all tissues tested (Fig. 4a, b). To determine the spatial expression patterns of *PEP1*, RNA *in situ* hybridization experiments were performed. Before vernalization, *PEP1* mRNA was strongly detected in the shoot apical meristem, the adjacent axillary meristems and young leaves (Fig. 4c). At the end of vernalization, *PEP1* mRNA could not be detected in the flower buds at the shoot apical meristem, and was only weakly expressed in the primordia of newly formed vegetative axillary shoots (Fig. 4d, f). In agreement with the qRT–PCR results, *PEP1* transcript levels were restored after vernalization in both flowering and vegetative meristems (Fig. 4e, g). Therefore, in contrast to *FLC* in *A. thaliana*, after vernalization *PEP1* mRNA is restored to similar levels to those present before vernalization.

In *A. thaliana* stable repression of *FLC* by vernalization correlates with modification of histone H3 (refs 8, 9), and particularly with the addition of three methyl groups to the lysine residue at position 27 (H3K27me3)[11,20–22]. During vernalization the level of this histone modification increases close to the transcription start site of *FLC*, and in growing tissues it persists after return to normal growth temperatures, spreading across the gene and correlating with stable transcriptional repression[11]. In *A. alpina* the level of H3K27me3 also increased at *PEP1* during vernalization, but its levels decreased again after vernalization (Fig. 4h). Therefore, in contrast to *FLC* in *A. thaliana*, in *A. alpina* the H3K27me3 mark accumulated at *PEP1* during vernalization but did not persist after vernalization.

Thus, temporal changes in *PEP1* expression contribute to the perennial life history of *A. alpina*. Repression of *PEP1* transcription during vernalization and reactivation after return to warm temperatures correlate with unstable modifications of H3K27me3 at the *PEP1* locus. This observation is in contrast to *FLC* expression patterns in growing tissues of *A. thaliana*, in which *FLC* expression is stably repressed by vernalization and the H3K27me3 chromatin mark persists at the locus after vernalization[10,11,20–22]. Annual plants such as *A. thaliana* only flower once in their lifetime, and therefore to maximise seed production all growing shoot meristems undergo the transition to flowering when the environmental conditions are optimal. Thus, stable repression of the floral repressor *FLC* by vernalization is consistent with the monocarpic life strategy of *A. thaliana*. Mutations in a component of the polycomb-group protein complex required to produce the H3K27me3 mark prevent stable repression of *FLC* by vernalization in *Arabidopsis*, indicating that this mark is required for epigenetic silencing of *FLC*[10,20]. Moreover, *Arabidopsis* accessions that show variation in the duration of vernalization required to saturate the flowering response show differences in the accumulation of the H3K27me3 mark at the *FLC* locus during vernalization[23]. However, a reduction in the level of the mark after vernalization, as we observed in apices of fully vernalized *A. alpina* plants, has not been observed in growing tissues of *A. thaliana*, indicating that this aspect of *FLC*/*PEP1* regulation differs between species. Nevertheless, in *A. thaliana* the H3K27me3 mark at *FLC* does not persist in cells of fully expanded leaves that are

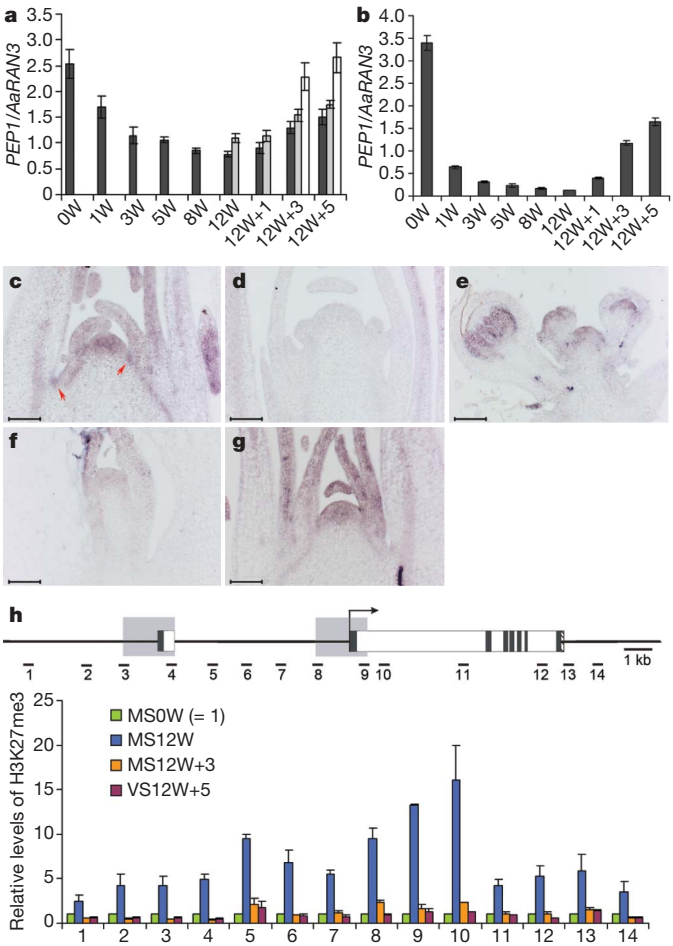


Figure 4 | Repression of *PEP1* expression by vernalization is unstable and correlated with changes in histone methylation. **a**, **b**, *PEP1* mRNA levels fall in apices (**a**) and leaves (**b**) during vernalization treatment (1 to 12 weeks (W)), and are restored after vernalization (12W + 1 to 5 weeks). RNA levels were measured by qRT–PCR. In **a**, dark grey bars represent main-shoot apices. Grey bars represent apices from axillary shoots produced before vernalization. *PEP1* expression in these axillary shoots was only measured at 12W, 12W + 1, 12W + 3 and 12W + 5 time points. White bars represent apices from axillary shoots produced after vernalization. In **a** and **b** the means of three technical replicates are shown; error bars indicate s.d. **c**–**e**, *In situ* hybridization of *PEP1* mRNA in the main-shoot apex after 8 weeks growth in long days before vernalization (**c**), followed by 12 weeks vernalization (**d**), and after a further 4 weeks in long days (**e**). Arrows highlight axillary meristems that will develop into vegetative axillary shoots during or after vernalization. **f**, **g**, *In situ* hybridization of *PEP1* mRNA in vegetative axillary shoots at the end of 12-weeks vernalization (**f**) and after a further 4 weeks in long days (**g**). Scale bars, 100 nm. **h**, The *AaFLC* locus showing the regions amplified by PCR after ChIP (top). Boxes represent exons (dark grey), introns (white), 3′ untranslated region (UTR; striped) and a 2-kb duplicated region present at the locus (grey). H3K27me3 levels determined by ChIP of 14 regions on the *AaFLC* locus (bottom). All data points were divided by the H3K27me3 level before vernalization (8 weeks in long days) to provide relative values. The level before vernalization is therefore arbitrarily assigned a value of 1.0. Main-shoot apices of *A. alpina* plants were grown for 8 weeks in long days (MS0W), after 12 weeks vernalization (MS12W), and followed by 3 weeks in long days (MS12W + 3). Apices from vegetative axillary shoots were grown for 5 weeks in long days after vernalization (VS12W + 5). The data presented are the means of two technical replicates, error bars denote s.e.m. Similar patterns of modification were obtained in a second set of biological replicates.

no longer undergoing mitosis, indicating that in these tissues the mechanism required for persistence of this mark at *FLC* is not active[11].

The instability of *PEP1* repression in *A. alpina* allows PEP1 to block flowering of all meristems that did not undergo flowering during the preceding vernalization treatment, and therefore to take on roles in the perennial life cycle that are not required in the annual life cycle of *A. thaliana*. However, although differential regulation of *PEP1* enhances perennial characters and confers seasonality, *pep1* mutants are still long-lived and behave as perennial plants. This indicates that the differential regulation of *PEP1* and *FLC* is not the only difference between *A. thaliana* and *A. alpina* that is responsible for perennialism. Furthermore, although our data demonstrate an important role for PEP1 as an inhibitor of flowering in the perennial cycle, the presence of extra components, including floral promoters that increase in expression before each flowering cycle, cannot be excluded. Nevertheless, our analysis of *PEP1* provides insight into the different mechanisms underlying seasonal flowering in perennial and annual plants consistent with their life history, and how gene functions involved in the control of seasonal flowering can be incorporated into perennial-specific traits such as polycarpy. The distinction between annual and perennial life histories has arisen independently many times in the flowering plants[24,25], and in the *Brassicaceae* both life histories occur in several genera[26], suggesting that they diversified independently. Furthermore, not all perennials flower in response to vernalization and FLC is not a universal regulator of vernalization requirement[27], supporting the idea that different mechanisms must regulate the perennial growth habit in other groups of plants. Species-specific traits can evolve through alterations in the expression patterns of regulatory genes[28]. The differences in histone modifications at *FLC* and *PEP1* in *A. thaliana* and *A. alpina*, respectively, suggest that differences in chromatin regulation may be one of the mechanisms by which these alterations in gene expression patterns occur, thereby allowing diversification of rapidly evolving traits such as life history characters.

METHODS SUMMARY

Plant material. *Arabis alpina* L. accession Pajares was collected in the Cordillera Cantábrica mountain system in Spain (42°59′32″ N, 5°45′32″ W; 1,400 m altitude) and self-fertilized for six generations by single-seed descent. *Arabidopsis thaliana* experiments were performed using the Columbia-0 accession.

Ethyl methanesulphonate mutagenesis. *Arabis alpina* seeds were mutagenized with 0.35% ethyl methanesulphonate (Sigma) for 8–9 h. Twenty-five-thousand M_2 seedlings (2,500 M_1 families) were screened in long days at a 20 °C controlled-environment glasshouse.

BAC analysis and sequencing. Two positive BAC clones were identified by screening an *A. alpina* BAC library made from accession Pajares (R. Castillo, unpublished data) using *A. thaliana FLC* as a probe without the sequence encoding the MADS box. The DNA sequence of a positively hybridizing BAC was determined using the Sanger method, assembled and annotated manually.

Plant transformation. Plasmid constructs were generated by cloning PCR-amplified fragments into GATEWAY compatible binary vectors and introduced into *Agrobacterium* strain GV3101 (pMP90RK). Transgenic *A. thaliana* and *A. alpina* were generated by the floral-dip method[29].

Gene expression analysis. Total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen) from expanded leaves, main-shoot apices and axillary-shoot apices grown before or after vernalization. cDNA synthesis and qRT–PCR were performed as previously described[30].

***In situ* hybridization.** *In situ* hybridization was performed on apices of the main shoot and of axillary shoots, as described previously[30].

Chromatin immunoprecipitation. Chromatin immunoprecipitation (ChIP) was carried out using apices of the main shoots (MS) of *A. alpina* plants grown for 8 weeks in long days (MS0W), then vernalized for 12 weeks (MS12W), and finally grown for 3 weeks in long days (MS12W + 3). Apices of axillary vegetative shoots grown during and after vernalization were collected after 5 weeks in long days (VS12W + 5). Chromatin samples were immunoprecipitated with anti-H3K27me3 (Upstate). Results were presented as a ratio of vernalized/non-vernalized samples and are the mean of two technical replicates.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 January; accepted 12 March 2009.
Published online 15 April 2009.

1. Battey, N. H. & Tooke, F. Molecular control and variation in the floral transition. *Curr. Opin. Plant Biol.* **5**, 62–68 (2002).

2. Turck, F., Fornara, F. & Coupland, G. Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Annu. Rev. Plant Biol.* **59,** 573–594 (2008).
3. Baurle, I. & Dean, C. The timing of developmental transitions in plants. *Cell* **125,** 655–664 (2006).
4. Ansell, S. W., Grundmann, M., Russell, S. J., Schneider, H. & Vogel, J. C. Genetic discontinuity, breeding-system change and population history of *Arabis alpina* in the Italian Peninsula and adjacent Alps. *Mol. Ecol.* **17,** 2245–2257 (2008).
5. Koch, M. A. *et al.* Three times out of Asia Minor: the phylogeography of *Arabis alpina* L. (*Brassicaceae*). *Mol. Ecol.* **15,** 825–839 (2006).
6. Michaels, S. D. & Amasino, R. M. *FLOWERING LOCUS C* encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* **11,** 949–956 (1999).
7. Sheldon, C. C. *et al.* The *FLF* MADS box gene: A repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell* **11,** 445–458 (1999).
8. Bastow, R. *et al.* Vernalization requires epigenetic silencing of *FLC* by histone methylation. *Nature* **427,** 164–167 (2004).
9. Sung, S. & Amasino, R. M. Vernalization in *Arabidopsis* thaliana is mediated by the PHD finger protein VIN3. *Nature* **427,** 159–164 (2004).
10. Gendall, A. R., Levy, Y. Y., Wilson, A. & Dean, C. The *VERNALIZATION 2* gene mediates the epigenetic regulation of vernalization in *Arabidopsis. Cell* **107,** 525–535 (2001).
11. Finnegan, E. J. & Dennis, E. S. Vernalization-induced trimethylation of histone H3 lysine 27 at *FLC* is not maintained in mitotically quiescent cells. *Curr. Biol.* **17,** 1978–1983 (2007).
12. Battey, N. H. Aspects of seasonality. *J. Exp. Bot.* **51,** 1769–1780 (2000).
13. Bohlenius, H. *et al.* CO/FT regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312,** 1040–1043 (2006).
14. Foster, T., Johnston, R. & Seleznyova, A. A morphological and quantitative characterization of early floral development in apple (*Malus* × *domestica* Borkh.). *Ann. Bot. (Lond.)* **92,** 199–206 (2003).
15. Diomaiuto, J. Periodic flowering or continual flowering as a function of temperature in a perennial species: the Ravenelle wallflower (*Cheiranthus cheiri* L.). *Phytomorphology* **38,** 163–171 (1988).
16. Hay, A. & Tsiantis, M. The genetic basis for differences in leaf form between *Arabidopsis thaliana* and its wild relative *Cardamine hirsuta. Nature Genet.* **38,** 942–947 (2006).
17. Hanikenne, M. *et al.* Evolution of metal hyperaccumulation required *cis*-regulatory changes and triplication of *HMA4. Nature* **453,** 391–395 (2008).
18. Schranz, M. E. *et al.* Characterization and effects of the replicated flowering time gene *FLC* in *Brassica rapa. Genetics* **162,** 1457–1468 (2002).
19. D'Aloia, M., Tocquin, P. & Perilleux, C. Vernalization-induced repression of *FLOWERING LOCUS C* stimulates flowering in *Sinapis alba* and enhances plant responsiveness to photoperiod. *New Phytol.* **178,** 755–765 (2008).
20. Schubert, D. *et al.* Silencing by plant Polycomb-group genes requires dispersed trimethylation of histone H3 at lysine 27. *EMBO J.* **25,** 4638–4649 (2006).
21. Sung, S., Schmitz, R. J. & Amasino, R. M. A. PHD finger protein involved in both the vernalization and photoperiod pathways in *Arabidopsis. Genes Dev.* **20,** 3244–3248 (2006).
22. Greb, T. *et al.* The PHD finger protein VRN5 functions in the epigenetic silencing of *Arabidopsis FLC. Curr. Biol.* **17,** 73–78 (2007).
23. Shindo, C., Lister, C., Crevillen, P., Nordborg, M. & Dean, C. Variation in the epigenetic silencing of *FLC* contributes to natural variation in *Arabidopsis* vernalization response. *Genes Dev.* **20,** 3079–3083 (2006).
24. Thomas, H., Thomas, H. M. & Ougham, H. Annuality, perenniality and cell death. *J. Exp. Bot.* **51,** 1781–1788 (2000).
25. Bena, G., Lejeune, B., Prosperi, J.-M. & Olivieri, I. Molecular phylogenetic approach for studying life-history evolution: the ambiguous example of the genus Medicago L. *Proc. R. Soc. Lond. B* **265,** 1141–1151 (1998).
26. Beilstein, M. A., Al-Shehbaz, I. A. & Kellogg, E. A. Brassicaceae phylogeny and trichome evolution. *Am. J. Bot.* **93,** 607–619 (2006).
27. Yan, L. *et al.* The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303,** 1640–1644 (2004).
28. Prud'homme, B., Gompel, N. & Carroll, S. B. Emerging principles of regulatory evolution. *Proc. Natl Acad. Sci. USA* **104,** 8605–8612 (2007).
29. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana. Plant J.* **16,** 735–743 (1998).
30. Searle, I. *et al.* The transcription factor FLC confers a flowering response to vernalization by repressing meristem competence and systemic signaling in *Arabidopsis. Genes Dev.* **20,** 898–912 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank P. Sharma for growing plants and R. Bruggmann for running the gene prediction pipeline. The laboratories of H.S. and G.C. are partly funded by a core grant from the Max Planck Society.

Author Information The GenBank accession number for the *PEP1* BAC sequence is FJ543377, and for the *PEP1* cDNA sequence is FJ755930. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.C. (coupland@mpiz-koeln.mpg.de) or M.C.A. (albani@mpiz-koeln.mpg.de).

METHODS

Growth conditions and phenotypic characterization. Plants were grown in long-day growth cabinets (16 h light at 22 °C, and 8 h dark at 18 °C) or long-day climate-controlled glasshouses (16 h light at 20°C), short-day cabinets (8 h light at 22 °C, and 16 h dark at 18 °C), and vernalized at 4 °C in short days (8 h light/16 h dark). To compare the duration of flowering between the *pep1*, *35S:AaFLC* dsRNAi line and wild-type plants, both genotypes were first grown for 5 weeks in long days (growth cabinet), vernalized for 12 weeks and then transferred back to long days. Flowering time was measured by scoring the leaves on the main shoot of at least eight individuals. Node position on the main shoot was determined by the presence of a true leaf and numbered from the bottom to the top of each plant.

Characterization of *pep1* mutant. Total RNA was extracted using an RNeasy Plant Mini Kit (Qiagen). Full-length *AaFLC* cDNA was amplified from expanded leaves and apices using primers on 5′ UTR (5′-AAACACAAAAAA AGAGTGAGAATAG-3′) and 3′ UTR (5′-AGTCTCTCAGCCATAGAGAG-3′) (30 cycles). PCR products were cloned into a pCR2.1-TOPO vector and 14 clones were sequenced. Genomic DNA from *pep1* mutant was extracted using the DNeasy DNA kit (Qiagen), and the *AaFLC* gene was sequenced using primers spanning the whole locus (primer sequences available on request). F_2 progenies were genotyped for the mutation in *AaFLC* using primers 5′-TTTGCCCCTTAG TTTTGTGG-3′ and 5′-TACCCGGGAAGACTACATGC-3′ flanking the mutation, and subsequently sequencing using primer 5′-TTTGCCCCTTAGTT TTGTGG-3′.

BAC analysis and sequencing. A BAC library was constructed using the pIndigoBAC-5 vector from partially HindIII-digested total genomic DNA of the accession Pajares (R. Castillo, unpublished data). The BAC library was screened using *Arabidopsis FLC* as a probe (without the MADS-box coding sequence) and two positive BACs were identified. Plasmid DNA was extracted from 100 ml overnight culture using the Qiagen Plasmid Midi kit. Positively hybridizing BACs were analysed by fingerprinting and all corresponded to one locus. The DNA sequence of a positively hybridizing BAC was determined using the Sanger method[31] and assembled using PHRED[32,33], PHRAP and Consed[34]. Gene prediction from GenMark.hmm[35], FGeneSH[36] and GenomeThreader[37] using several plant expressed-sequence-tag databases were integrated manually using the APOLLO genome editor[38]. Manual functional annotation of all predicted genes was on the basis of their homology to *A. thaliana* and other plant species, as well as on the presence of conserved domains identified using the InterPro[39] database.

Plasmid constructs. To generate the *35S::AaFLC* construct, the full-length (ATG to TAG) coding sequence of *AaFLC* was PCR-amplified with primer pair X-*AaFLC5*-F, 5′-(attB1)GAGACAGAAGCCATGGGTAG-3′, and X-*AaFLC3*-R, 5′-(attB2)GGCTTAATTGAGTAGTGGGAG-3′, and verified by sequencing after cloning into pDONR221 (Invitrogen). Through site-directed recombination it was subsequently cloned into a binary vector pLEELA, which is a derivative of pJawohl3 RNAi (GenBank accession AF404854) containing a GATEWAY cassette introduced into the HpaI site.

To make the AaFLC dsRNAi construct for transformation of *A. alpina*, a fragment including sequences from exon 4 to 3′ UTR was amplified from cDNA using primer pair *AaFLC*-RNAi-F1, 5′-(attB1)AAGCTCGTGGCA CCAAATGTC-3′, and *AaFLC*-RNAi-R1, 5′-(attB1)AGTCTCTCAGCCATAGA GAGT-3′. The sequence of the *AaFLC* RNAi fragment is between coordinates 364 and 763 in the *AaFLC/PEP1* cDNA sequence (GenBank Accession Number FJ755930). The fragment was recombined to generate a GATEWAY entry clone and subsequently recombined into the binary vector pJawohl8 (GenBank accession AF408413) to generate an intron-spliced hairpin construct.

Gene expression analysis. Total RNA was extracted from expanded leaves and shoot apices (approximately 1 mm in length), and an on-column DNase treatment (Qiagen) was performed to reduce any DNA contamination. RNA was analysed by qRT–PCR using a BioRad iQ5 apparatus and SYBR Green I detection. Three micrograms of total RNA was used for cDNA synthesis with oligo-dT15 as a primer. cDNA was diluted to a final volume of 180 µl, and 3.5 µl of diluted cDNA was used for PCR. An *A. alpina RAN3* gene was used as a control to normalize the amounts of cDNA. For testing *AaFLC/PEP1* expression with qRT–PCR, the primer pair *PEP1*-RT-F1, 5′-CTTGTCGTCTCCTCCTCTGG-3′, and *PEP1*-RT-R1, 5′-ACTACGGCGAGAGCAGTTTC-3′ were used. Primer pair *AaRAN3*-F2, 5′-CACAGGAAAAACCACATTCGT-3′, and *AaRAN3*-R2, 5′-CCATCCCTAAGACCACCAAAT-3′ were used to detect *AaRAN3* transcript abundance.

***In situ* hybridization.** Methods of digoxigenin labelling of mRNA probes, tissue preparation and *in situ* hybridization were performed as already described[30]. Template DNA used for probe synthesis to detect *AaFLC/PEP1* transcripts were PCR-amplified from cDNA using the primer pair T3-*PEP1*-F2, 5′-ATTA ACCCTCACTAAAGGGAAGCCAGATGGAGAAGAAGAC-3′, and T7-*PEP1*-R2, 5′-TAATACGACTCACTATAGGGAACAAGGGTACGAAGATCCA-3′. The underlined nucleotide sequence indicates either the T3 or the T7 RNA polymerase-binding sites.

Chromatin immunoprecipitation. ChIP were carried out as described[30]. The antibody anti-H3K27me3 (Upstate) was used for the immunoprecipitation and an anti-rat IgG (Sigma) was used as a negative control. Two independent immunoprecipitations were performed with the anti-H3K27me3 antibody as technical replicates, and average values were obtained. The ChIP DNA was subjected to 33–34 cycles of PCR (non-saturated conditions). Primers used for PCR were designed to avoid the duplicated region in the *AaFLC* locus represented in Fig. 4h (Supplementary Table 2). *RAN3* amplification was used to normalize DNA concentrations. The semi-quantification of the data was performed using the QIAxcel System (Qiagen) and the QX DNA size marker 50–800-bp (Qiagen). An appropriate dilution of the input amplified in similar conditions was used to compare the ChIP data. Results are presented as a ratio of vernalized/non-vernalized samples and are the means of two technical replicates. Similar patterns of modification were obtained in a second set of biological replicates.

Scanning electron microscopy. Plants of similar sizes were selected and marked on the main shoot just below node 15 (as in Supplementary Fig. 2). Apices from the main shoot were sampled to include leaves grown above node 15 at: 8 weeks in long days, after 12 weeks in vernalization and after 1 subsequent week in long days. Expanded leaves were removed from shoots apices and fixed in 2.5% glutaraldehyde in 0.002 M phosphate buffer (pH 7) and 0.01% Nonidet P-40. Samples were dehydrated in ethanol and critical-point dried in liquid CO_2. Leaf primordia were removed under a binocular microscope and samples were mounted on stubs, coated and subjected to high resolution Zeis Supra 55 VP FEG scanning electron microscopy with a Gatan Alto 2500 cryo system. The presence of flowers or flowering shoots and the identity of shoots in the axils of each leaf were recorded starting from the top of each sample until the lower leaf, which corresponded to the leaf in node 15. Each plant was scored at each developmental stage. At every node, the presence or absence of a shoot was recorded, shoots were excised and leaf primordia were removed under the binocular microscope and subjected to high resolution scanning electron microscopy.

Phylogenetic analysis. Protein–protein BLAST (http://www.ncbi.nlm.nih.gov/BLAST) was performed using the deduced amino acid sequence from AaFLC. ClustalX was used to create multiple alignments[40], and a phylogenetic tree was generated by the neighbour-joining method with Splitstree[41]. Bootstrap analysis was performed to estimate nodal support on the basis of 1,000 replicates. The accession numbers were: FLC, NP_196576; MAF1, NP_177833; MAF2, Q9FPN7; MAF3, NP_201311; MAF4, NP_201312; MAF5, AAO65320; AlFLC, AAV51231; ThFLC, AAY34137; AsFLC, AAZ92553; ArFLC, AAZ92552; BnFLC1, AAK70215; BnFLC2, AAK70216; BnFLC3, AAK70217; BnFLC4, AAK70218; BnFLC5, AAK70219; BrFLC1, ABI29999; BrFLC2, ABO40820; BrFLC3, ABI30001; BoFLC1, CAJ77613; BoFLC3, AAP31677; BoFLC4, AAQ76273; SaFLC, ABP96967; RsFLC, AAP31676; and AGL42, NP_568952.

31. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).
32. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
33. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
34. Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
35. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
36. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10,** 516–522 (2000).
37. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47,** 965–978 (2005).
38. Lewis, S. E. *et al.* Apollo: a sequence annotation editor. *Genome Biol.* **3,** research0082 (2002).
39. Apweiler, R. *et al.* InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16,** 1145–1150 (2000).
40. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882 (1997).
41. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14,** 68–73 (1998).

Vol 459|21 May 2009|doi:10.1038/nature08012

LETTERS

Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis

Sabrina L. Spencer[1,2]*, Suzanne Gaudet[1]†*, John G. Albeck[1], John M. Burke[1] & Peter K. Sorger[1]

In microorganisms, noise in gene expression gives rise to cell-to-cell variability in protein concentrations[1–7]. In mammalian cells, protein levels also vary[8–10] and individual cells differ widely in their responsiveness to uniform physiological stimuli[11–15]. In the case of apoptosis mediated by TRAIL (tumour necrosis factor (TNF)-related apoptosis-inducing ligand) it is common for some cells in a clonal population to die while others survive—a striking divergence in cell fate. Among cells that die, the time between TRAIL exposure and caspase activation is highly variable. Here we image sister cells expressing reporters of caspase activation and mitochondrial outer membrane permeabilization after exposure to TRAIL. We show that naturally occurring differences in the levels or states of proteins regulating receptor-mediated apoptosis are the primary causes of cell-to-cell variability in the timing and probability of death in human cell lines. Protein state is transmitted from mother to daughter, giving rise to transient heritability in fate, but protein synthesis promotes rapid divergence so that sister cells soon become no more similar to each other than pairs of cells chosen at random. Our results have implications for understanding 'fractional killing' of tumour cells after exposure to chemotherapy, and for variability in mammalian signal transduction in general.

TRAIL elicits a heterogeneous phenotypic response in both sensitive and relatively resistant cell lines: some cells die within 45 min, others 8–12 h later, and yet others live indefinitely (Supplementary Fig. 1). During the variable delay between TRAIL addition and mitochondrial outer membrane permeabilization (MOMP), upstream initiator caspases are active but downstream effector caspases are not[11,12]. Possible sources of cell-to-cell variability in response to TRAIL include genetic or epigenetic differences, stochastic fluctuations in biochemical reactions involving low copy number components ('intrinsic noise'[3]), differences in cell cycle phase, and natural variation in the concentrations of important reactants. To distinguish between these and other possibilities, we used live-cell microscopy to compare the timing and probability of death in sister cells exposed to TRAIL. If phenotypic variability is caused by genetic or epigenetic differences, sister cells should behave identically. In contrast, if stochastic fluctuations in reactions triggered by TRAIL predominate, sister cells should be no more similar to each other than pairs of cells selected at random. The influence of cell cycle state on apoptosis should be readily observable from time-lapse imaging of asynchronous cultures. Furthermore, variability arising from differences in protein levels (or in their activity or modification state) should produce a highly distinctive form of inheritance in which newly born sister cells are very similar, because they inherit similar numbers of abundant factors from their mother[4,7], but then diverge as new proteins are made and levels drift[10,16]. With this in mind, we examined apoptosis in HeLa cells and in non-transformed MCF10A mammary epithelial cells in the presence and absence of protein synthesis inhibitors.

Pairs of sister cells expressing a fluorescent reporter of MOMP (mitochondrial intermembrane space reporter protein, IMS-RP[11]) born during a 20–30 h period were identified by time-lapse microscopy. TRAIL and the protein synthesis inhibitor cycloheximide were then added and imaging continued for another 8 h. The TRAIL-to-MOMP interval (T_d) was calculated for each cell (Fig. 1a). Among the recently divided sisters (<7 h between division and death), T_d was highly correlated ($R^2 = 0.93$, Fig. 1b), whereas T_d was uncorrelated ($R^2 = 0.04$) for recently divided cells chosen at random. The time since division (Fig. 1c) and the position in the dish (data not shown) did not correlate with T_d, ruling out a role for cell cycle state and cell–cell interactions under our experimental conditions. However, as the time since division increased, sister-to-sister correlation in T_d decayed exponentially with a half-life of ~11 h, so that sisters lost memory of shared ancestry within ~50 h or about two cell generations ($R^2 \leq 0.05$, the same as random pairs of cells; Fig. 1d, e). Similar results were obtained with MCF10A cells (Supplementary Fig. 2).

High correlation between recently born sisters shows that the variability in T_d arises from differences that exist before TRAIL exposure, and rules out stochastic fluctuations in signalling reactions. Rapid decorrelation also rules out genetic mutation or conventional epigenetic differences (which typically last $10–10^5$ cell divisions[17]). However, transient heritability is precisely what we expected for cell-to-cell differences arising from variations in the concentrations or states of proteins that are partitioned binomially at cell division.

Whereas all TRAIL-treated HeLa cells eventually died in the presence of cycloheximide, in its absence a fraction always survived (presumably owing to induction of survival pathways[18]). When the fates of sister cells were compared, both lived or both died in most cases (chi-squared test, $P = 7 \times 10^{-19}$, Supplementary Fig. 3). Variability in T_d across the population was large (Fig. 1g and Supplementary Fig. 4), but recently born sisters were nevertheless correlated in T_d ($R^2 = 0.75$, Fig. 1f). Again, cell cycle phase was not correlated with fate or the time-to-death (Fig. 1g). Decorrelation in T_d among sisters was an order of magnitude more rapid in the presence of protein synthesis than in its absence (~1.5 h half-life, Fig. 1h, i and Supplementary Fig. 5). Thus, the length of time that T_d is heritable is very sensitive to rates of protein synthesis, both basal and TRAIL-induced.

We then asked whether the concentrations of proteins regulating TRAIL-induced apoptosis are sufficiently different from cell to cell to account for the variability in T_d. Using flow cytometry, we measured the distributions of five apoptotic regulators for which specific antibodies are available. All five proteins were log-normally distributed across the population with coefficients of variation ranging between 0.21 and 0.28 for cells of similar size (Fig. 2a), consistent with data on

[1]Center for Cell Decision Processes, Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. †Present address: Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

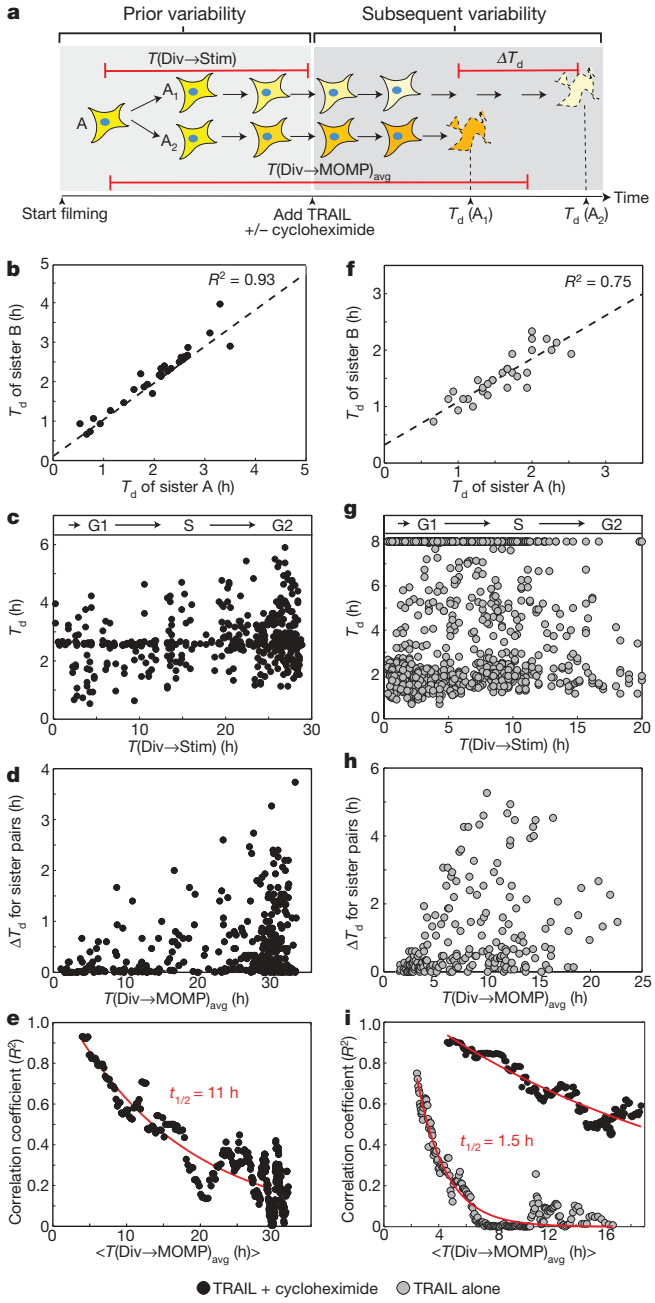
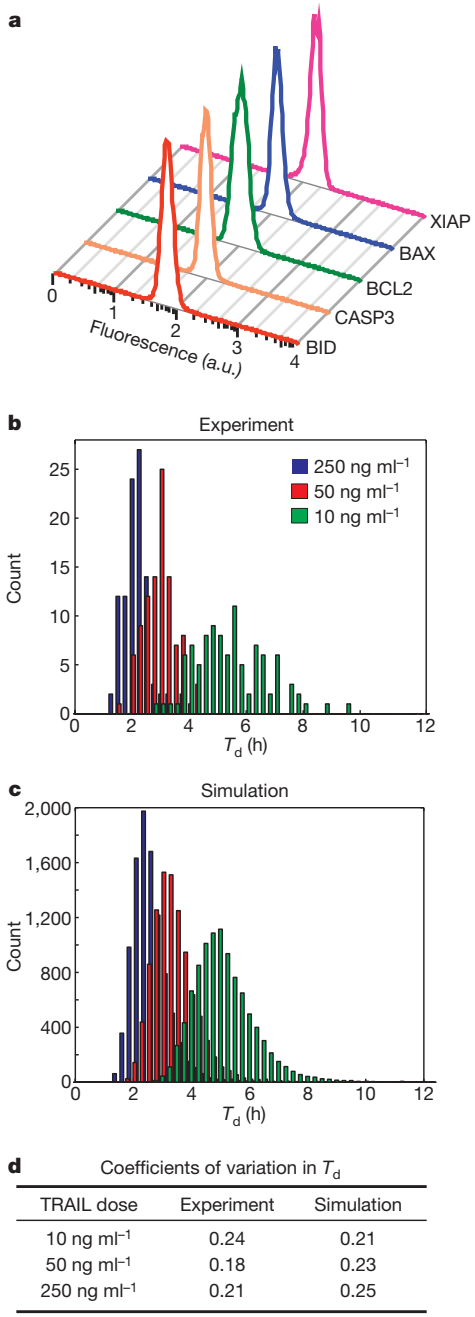


Figure 1 | The time-to-death is highly correlated between HeLa sister cells, but correlation decays as a function of time since division. **a**, Schematic of the experimental design. ΔT_d represents the difference in time of MOMP between sisters; $T(\text{Div}\rightarrow\text{MOMP})_{avg}$ denotes the time between cytokinesis of the mother and the average time of MOMP in daughter cells; $T(\text{Div}\rightarrow\text{Stim})$ denotes the time between cytokinesis and TRAIL addition. The shading of each cell depicts concentrations/states of relevant proteins. **b**, **f**, Similarity in T_d among pairs of recently divided sister cells ($T(\text{Div}\rightarrow\text{MOMP})_{avg} < 7$ h for **b** and < 3.5 h for **f**). **c**, **g**, T_d as a function of $T(\text{Div}\rightarrow\text{Stim})$, a proxy for cell cycle state ($R^2 < 0.03$). **d**, **h**, ΔT_d as a function of $T(\text{Div}\rightarrow\text{MOMP})_{avg}$. **e**, **i**, Decay in the correlation of T_d between sister pairs as a function of $T(\text{Div}\rightarrow\text{MOMP})_{avg}$. In **i**, black circles represent data for cells treated with TRAIL plus cycloheximide, imaged under the same conditions as the TRAIL-alone treatment (Supplementary Fig. 5).



d Coefficients of variation in T_d

TRAIL dose	Experiment	Simulation
10 ng ml^{-1}	0.24	0.21
50 ng ml^{-1}	0.18	0.23
250 ng ml^{-1}	0.21	0.25

Figure 2 | Endogenous variation in the concentrations of apoptotic regulators is sufficient to explain variability in T_d. **a**, Protein distributions in untreated HeLa cells determined by flow cytometry. a.u., arbitrary units. **b**, **c**, Distributions of T_d for HeLa cells treated with TRAIL at concentrations indicated (with cycloheximide) as determined experimentally (**b**) or estimated by simulations (**c**). **d**, Coefficients of variation for distributions in **b** and **c**.

other proteins[10]. To determine the impact of variability in protein levels on variability in time-to-death, we turned to an ordinary differential equation model of TRAIL-induced apoptosis[12]. This model encapsulates the biochemistry of TRAIL-mediated death and recapitulates the dynamics of apoptosis under various conditions of protein depletion or overexpression[12]. When variability in T_d arising from variance in protein levels was modelled, a good match was observed to experimental data (Fig. 2b–d) indicating that the measured differences in protein levels are sufficient to account for variability in T_d.

Next, we investigated which steps in receptor-mediated apoptosis are responsible for variation in time-to-death. To address this question, we grouped reactions into three sets: those occurring before, during, or subsequent to MOMP (Fig. 3a, blue, grey and orange). Before MOMP, TRAIL binds and oligomerizes DR4/5 receptors, promoting assembly of death-inducing signalling complexes (DISCs) that then activate initiator pro-caspases-8 and -10 (CASP8 and CASP10)[19]. Active CASP8/10 cleave BID to a truncated form

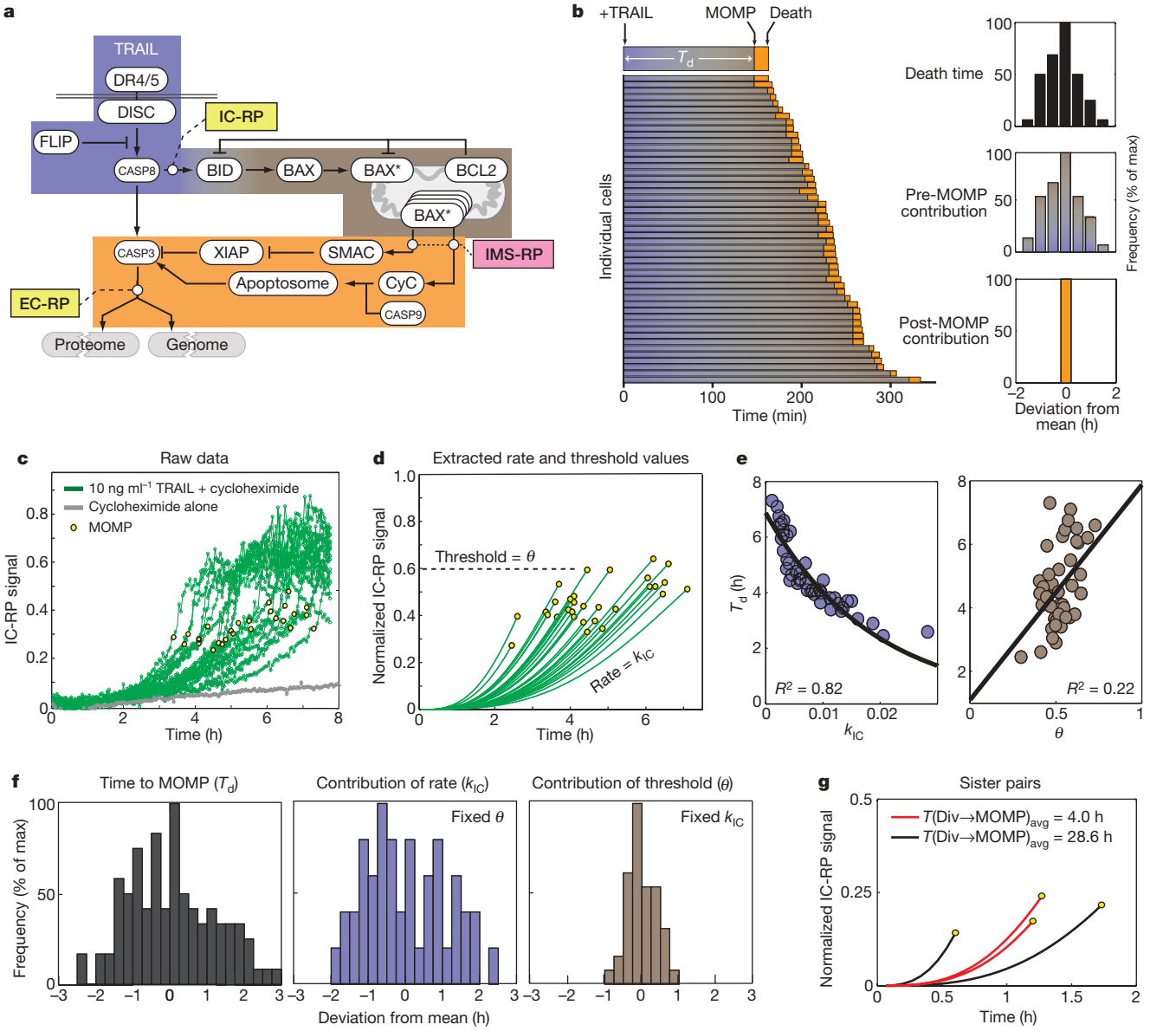


Figure 3 | A single time-dependent process upstream of MOMP predicts the time-to-death. **a**, Schematic of receptor-mediated apoptosis signalling, with IC-RP, EC-RP and IMS-RP indicated. The BCL2 protein family is represented in simplified form by BID, BAX and BCL2. Reactions occur before (blue), during (grey), or subsequent to MOMP (orange). BAX*, activated BAX; CyC, cytochrome *c*. **b**, Timing of apoptotic events in HeLa cells expressing IMS-RP and EC-RP and treated with 50 ng ml^{-1} TRAIL plus cycloheximide; blue-grey denotes the pre-MOMP interval, and orange denotes the interval between MOMP and half-maximal cleavage of EC-RP (a marker of death). Insets show death times computed from data (top) and contributions of pre-MOMP (middle) or post-MOMP (bottom) intervals. **c**, **d**, Raw (**c**) and fitted (**d**) trajectories for IC-RP cleavage in single TRAIL-treated HeLa cells co-expressing IMS-RP and IC-RP. Values for height of the MOMP threshold (θ) and rate of approach to the threshold (k_{IC}) were derived by fitting (Supplementary Fig. 6). **e**, Correlation between T_d and either k_{IC} (left) or θ (right), for data in **d**. **f**, Relative contributions of variability in k_{IC} (blue) or θ (grey) to variability in T_d (black; Supplementary Fig. 7). **g**, Representative trajectories of IC-RP cleavage in recently divided sister HeLa cells having similar T_d (red) and older sisters with differing T_d (black), treated with 50 ng ml^{-1} TRAIL plus cycloheximide.

(tBID)[20,21], which then activates the pore-forming proteins BAX and BAK[22]. CASP8 and CASP10 also process effector pro-caspases-3 and -7 (CASP3 and CASP7), but effector caspase activity is held in check by XIAP until MOMP[19]. MOMP itself involves self-assembly of activated BAX and BAK into transmembrane pores, a process antagonized by anti-apoptotic BCL2 family proteins[22]. When levels of activated tBID, BAX and BAK exceed a threshold set by inhibitory BCL2 proteins, pores form in the mitochondrial outer membrane, allowing cytochrome *c* and SMAC (also known as DIABLO) to translocate into the cytosol[22]. In post-MOMP reactions, cytosolic SMAC neutralizes XIAP, relieving CASP3 and CASP7 inhibition, and allowing cleavage of effector caspase substrates and consequent cell death[19]. In a parallel route to CASP3 and CASP7 activation, cytosolic cytochrome *c* promotes apoptosome assembly and caspase-9 activation.

To establish which steps in TRAIL-induced apoptosis have the greatest impact in determining variability in death time, we imaged cells expressing a reporter protein of either initiator or effector caspase activity (IC-RP or EC-RP)[11] in combination with IMS-RP. We found almost all variability in T_d to arise during the pre-MOMP interval (Fig. 3b). The timing of MOMP itself is determined by the rate at which tBID accumulates to a threshold set by the levels of BCL2 family proteins. This rate and threshold can be inferred from the initial rate of IC-RP cleavage (k_{IC}) and the fraction of IC-RP cleaved (θ) at the time of MOMP, respectively. When k_{IC} and θ were measured in single TRAIL-treated cells, the timing of MOMP was found to be controlled by a variable rate of approach to a threshold of variable height (Fig. 3c, d). However, variation in k_{IC} had a significantly greater role in determining T_d than variation in θ ($R^2 = 0.82$ versus $R^2 = 0.22$;

Fig. 3e, f and Supplementary Fig. 7). Moreover, k_{IC} was very similar in recently born sister cells with similar T_d, but dissimilar in older sisters (Fig. 3g). We conclude that cell-to-cell variability in k_{IC}—and by implication the rate of conversion of BID to tBID—is the primary determinant of variability in the time-to-death under our experimental conditions.

Levels of several proteins set k_{IC}, including DR4/5 receptors, DISC components, CASP8 and BID itself. Modelling suggested that knowing the concentration of any single protein upstream of BID would have minimal value in predicting T_d—the impact of variation in all other proteins is too great (Fig. 4a). Live-cell analysis of FLIP (also known as CFLAR), an important regulator of pro-caspase-8 binding to the DISC, was consistent with this prediction, as was analysis of other single proteins by flow cytometry (Fig. 4b and data not shown). However, modelling showed that with increasing overproduction of BID, measurement of its levels would be increasingly predictive of T_d (Fig. 4c, Supplementary Fig. 8). We therefore measured the relationship between dispersion in T_d and levels of BID tagged with green-fluorescent-protein (GFP) (Fig. 4d). A ~50-fold increase in BID caused the variability in T_d to fall significantly, concomitant with a decrease in mean time-to-death from ~3 h to ~45 min. Thus, only when overexpressed is the level of one protein predictive of T_d; under normal circumstances, control of T_d is multivariate.

Other studies (for example, ref. 23.) address genetic factors determining the average sensitivity of cell lines to TRAIL, whereas this paper examines non-genetic sources of cell-to-cell variability within an individual cell line. We come to three primary conclusions. First, cell-to-cell variation in the timing and probability of death is transiently heritable. Cell cycle state, the number of neighbouring cells, and stochastic fluctuations in TRAIL-induced signalling reactions do not have a crucial involvement under our conditions. Instead, variability in phenotype arises from cell-to-cell differences in protein levels that exist before TRAIL exposure (our experiments do not distinguish between cell-to-cell differences in total concentrations and in post-translationally modified forms). Second, the rate at which sisters lose memory of a shared past is an order of magnitude faster in the presence of protein translation than in its absence. This further implicates variability in protein levels as the origin of differences in phenotype. Third, knowing the concentration of individual proteins does not allow T_d to be predicted, but measuring the rate of a single reaction does (BID to tBID conversion in our experiments). This finding probably holds for other examples of ligand-induced apoptosis; however for intrinsic apoptosis, different proteins will control the rate of approach to MOMP and θ may dominate in certain contexts. The fundamental point is that each of these properties is determined by the levels and activities of several proteins. Given the prevalence of multi-protein cascades in signal transduction, multivariate control over cell-to-cell variability is likely to be more common than the univariate control observed in other settings[8,23,24].

Heritable, non-genetic determinants of phenotype are often referred to as 'epigenetic'[17], but the transient heritability we observe here is fundamentally different in origin and duration. Given variability in growth rates and noise in gene expression, genetically identical cells will inevitably contain slightly different concentrations of most proteins. However, differences in protein concentrations do not necessarily affect phenotype, a property often referred to as robustness[25]. For example, the efficiency with which effector caspase substrates are cleaved does not vary much from cell to cell[12]. Given the importance of tight control over apoptosis, cell-to-cell variability in the timing and probability of death seems unlikely to reflect an inability of cells to achieve robust regulation. Instead, by transforming what is a binary decision at the single-cell level into a graded response at the population level, variability probably has an adaptive advantage.

TRAIL is at present undergoing clinical trials as an anti-cancer drug[26] and our findings may have implications for the use of TRAIL and other apoptosis inducers as therapeutics. Many drugs exhibit 'fractional killing', in which each round of therapy kills some but not all of the cells in a tumour[27]. Traditionally, this is thought to reflect differences in genotype, cell cycle state, or the involvement of

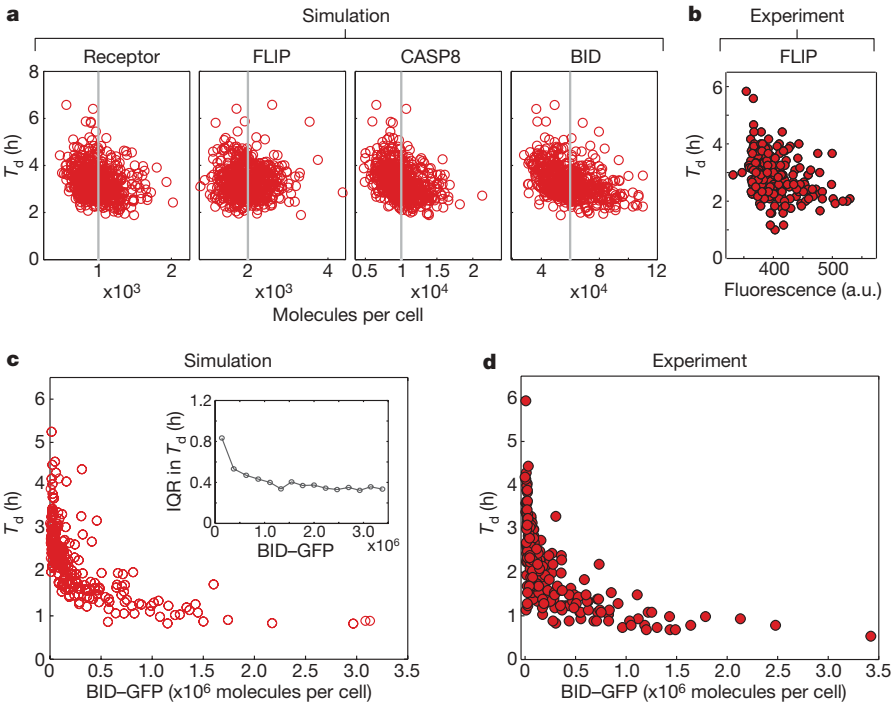


Figure 4 | No single protein predicts T_d under normal conditions but overexpression can increase predictability. **a**, T_d as a function of four protein levels, based on simulation as in Fig. 2c. Grey lines denote mean protein concentration; each point represents a single simulated cell. **b**, Death time as a function of endogenous FLIP levels in H1299 cells. **c**, **d**, Effect of BID–GFP overexpression on T_d in HeLa cells, as predicted by simulation (**c**) or observed in experiment (**d**). Inset shows the reduction in dispersion of T_d with increasing BID–GFP, as measured by the interquartile range (IQR; Supplementary Fig. 8).

cancer stem cells, but our data demonstrate that marked variability can also arise from natural differences in protein levels. We propose that the efficiency of TRAIL-mediated killing of cancer cells could be increased by reducing the impact of cell-to-cell variability, perhaps through co-drugging.

METHODS SUMMARY

Live-cell microscopy. Cells expressing IMS-RP and Förster resonance energy transfer (FRET) reporters EC-RP or IC-RP were imaged as described[11]. In Fig. 1, cells were imaged for 20–30 h to determine the time of division and to identify sister pairs; media containing 50 ng ml^{-1} TRAIL plus 2.5 μg ml^{-1} cycloheximide, or 250 ng ml^{-1} TRAIL alone, was then added. The difference in TRAIL concentrations was designed to generate a similar range in T_d with and without cycloheximide (Supplementary Fig. 4). Cells were then imaged for 8 h to determine the time of MOMP, by monitoring cytosolic translocation of IMS-RP. Unless otherwise noted, all treatments included 2.5 μg ml^{-1} cycloheximide.

Data analysis. Correlation coefficients (R^2) were obtained by linear regression except where noted. Sister–sister correlation was determined by sorting pairs of cells on $T(\mathrm{Div}\rightarrow\mathrm{MOMP})_{\mathrm{avg}}$ (where 'Div' denotes division) and calculating R^2 for the first 40 pairs. R^2 was then recalculated for cells 2–41, 3–42, and so on, and the results were plotted as a function of the average $T(\mathrm{Div}\rightarrow\mathrm{MOMP})_{\mathrm{avg}}$ for the 40 cells in question, denoted by '$<\ >$'. The results were fit to an exponential decay $R^2 = 1.2e^{(-0.063\ T(\mathrm{Div}\rightarrow\mathrm{MOMP})_{\mathrm{avg}})}$ for TRAIL plus cycloheximide, and $R^2 = 2.3e^{(-0.47T(\mathrm{Div}\rightarrow\mathrm{MOMP})_{\mathrm{avg}})}$ for TRAIL alone. Half-lives were calculated as $\ln(2)/0.063 = 11$ h, and $\ln(2)/0.47 = 1.5$ h. Contributions to T_d of k_{IC}, θ and pre- and post-MOMP intervals were obtained by fixing one parameter at the mean value and allowing the other to vary over the observed range, then mean-centring the resulting distributions (Supplementary Fig. 7). Fitted IC-RP trajectories were obtained after subtracting a trajectory for cycloheximide alone (Fig. 3c) to control for photobleaching (Supplementary Fig. 6).

Modelling. The responses of cell populations were simulated using a trained ordinary differential equation model[12] sampling from log-normally distributed protein concentrations with coefficient of variation (CV) ≈ 0.25 (see Methods). In Fig. 4, BID–GFP (an experimental observable) was added to log-normally distributed endogenous BID (unobservable); other proteins were sampled from log-normal distributions as before. Simulations were adjusted to match the distribution BID–GFP achieved experimentally.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 March; accepted 25 March 2009.
Published online 12 April 2009.

1. Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422,** 633–637 (2003).
2. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437,** 699–706 (2005).
3. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).
4. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual vacteria *Cell* **123,** 1025–1036 (2005).
5. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94,** 814–819 (1997).
6. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31,** 69–73 (2002).
7. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307,** 1962–1965 (2005).
8. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453,** 544–547 (2008).
9. Feinerman, O., Veiga, J., Dorfman, J. R., Germain, R. N. & Altan-Bonnet, G. Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science*, **321,** 1081–1084 (2008).
10. Sigal, A. *et al.* Variability and memory of protein levels in human cells. *Nature* **444,** 643–646 (2006).
11. Albeck, J. G. *et al.* Quantitative analysis of pathways controlling extrinsic apoptosis in single cells. *Mol. Cell* **30,** 11–25 (2008).
12. Albeck, J. G., Burke, J. M., Spencer, S. L., Lauffenburger, D. A. & Sorger, P. K. Modeling a snap-action, variable-delay switch controlling extrinsic cell death. *PLoS Biol.* **6,** 2831 (2008).
13. Geva-Zatorsky, N. *et al.* Oscillations and variability in the p53 system. *Mol. Syst. Biol.* **2,** 2006.0033 (2006).
14. Goldstein, J. C., Kluck, R. M. & Green, D. R. A single cell analysis of apoptosis. Ordering the apoptotic phenotype. *Ann. NY Acad. Sci.* **926,** 132–141 (2000).
15. Lahav, G. *et al.* Dynamics of the p53–Mdm2 feedback loop in individual cells. *Nature Genet.* **36,** 147–150 (2004).
16. Kaufmann, B. B., Yang, Q., Mettetal, J. T. & van Oudenaarden, A. Heritable stochastic switching revealed by single-cell genealogy. *PLoS Biol.* **5,** e239 (2007).
17. Rando, O. J. & Verstrepen, K. J. Timescales of genetic and epigenetic inheritance. *Cell* **128,** 655–668 (2007).
18. Chaudhary, P. M., Eby, M., Jasmin, A., Bookwalter, A., Murray, J. & Hood, L. Death receptor 5, a new member of the TNFR family, and DR4 induce FADD-dependent apoptosis and activate the NF-κB pathway. *Immunity* **7,** 821–830 (1997).
19. Fuentes-Prior, P. & Salvesen, G. S. The protein structures that shape caspase activity, specificity, activation and inhibition. *Biochem. J.* **384,** 201–232 (2004).
20. Li, H., Zhu, H., Xu, C. & Yuan, J. Cleavage of BID by caspase 8 mediates the mitochondrial damage in the Fas pathway of apoptosis. *Cell* **94,** 491–501 (1998).
21. Luo, X., Budihardjo, I., Zou, H., Slaughter, C. & Wang, X. Bid, a Bcl2 interacting protein, mediates cytochrome c release from mitochondria in response to activation of cell surface death receptors. *Cell* **94,** 481–490 (1998).
22. Youle, R. J. & Strasser, A. The BCL-2 protein family: opposing activities that mediate cell death. *Nature Rev. Mol. Cell Biol.* **9,** 47–59 (2008).
23. Wagner, K. W. *et al.* Death-receptor O-glycosylation controls tumor-cell sensitivity to the proapoptotic ligand Apo2L/TRAIL. *Nature Med.* **13,** 1070–1077 (2007).
24. Cohen, A. A. *et al.* Dynamic proteomics of individual cancer cells in response to a drug. *Science*, **322,** 1511–1516 (2008).
25. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387,** 913–917 (1997).
26. Ashkenazi, A. & Herbst, R. S. To kill a tumor cell: the potential of proapoptotic receptor agonists. *J. Clin. Invest.* **118,** 1979–1990 (2008).
27. Berenbaum, M. C. *In vivo* determination of the fractional kill of human tumor cells by chemotherapeutic agents. *Cancer Chemother. Rep.* **56,** 563–571 (1972).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Flusberg, S. Govind, L. Kleiman, A. Letai, B. Millard, R. Milo, T. Norman, J. Paulsson and R. Ward for their help. This work was supported by National Institute of Health (NIH grants) GM68762 and CA112967.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.K.S. (peter_sorger@hms.harvard.edu).

METHODS

Cell culture and transfections. HeLa cells were maintained in DMEM (Mediatech, Inc.) supplemented with L-glutamine (Gibco), penicillin/streptomycin (Gibco) and 10% fetal bovine serum (FBS; Mediatech, Inc.). MCF10A cells were cultured as described[28]. H1299 cells containing FLIP tagged with yellow fluorescent protein (YFP) at the endogenous locus were obtained from the Kahn Dynamic Proteomics Project and maintained at 8% CO_2 in RPMI (Mediatech, Inc.) with 10% FBS and penicillin/streptomycin. HeLa cells expressing IMS-RP[11] were transfected using FuGENE 6 (Roche) with pd4-BID-EGFP (Clontech) to sample expression levels across a wide range. H1666 cells were maintained (and imaged) in RPMI supplemented with 10% FBS, L-glutamine, penicillin/streptomycin, 1× ITES (Lonza Biosciences), 50 nM hydrocortisone, 10 μM phosphorylethanolamine, 0.1 nM Tri-iodothyronine, 10 mM HEPES, 0.5 mM sodium pyruvate, 2 g l^{-1} BSA and 1 ng ml^{-1} EGF. Fresh primary liver cells were obtained from CellzDirect and plated on collagen type I. Cells were maintained (and imaged) in Eagle's Minimum Essential Medium supplemented with 10% FBS, L-glutamine, 100 nM dexamethasone, 5 μg ml^{-1} human insulin, 5 μg ml^{-1} transferrin from human serum, 5 μg ml^{-1} sodium selenite and 15 mM HEPES. SKBR3 cells were maintained (and imaged) in RPMI supplemented with 10% FBS, L-glutamine and penicillin/streptomycin.

Live-cell microscopy. HeLa cells expressing IMS-RP and FRET reporters EC-RP or IC-RP were imaged in a 37 °C humidified chamber as described[11]. For sister cell experiments, HeLa cells expressing IMS-RP were imaged[11] at 10-min intervals for 20–30 h in phenol-red-free CO_2-independent medium (Invitrogen) with L-glutamine, penicillin/streptomycin and 1% serum (Fig. 1b–e), or at ~5% CO_2 in phenol-red-free DMEM with L-glutamine, penicillin/streptomycin and 10% serum (Fig. 1f–i, see also Supplementary Fig. 5). The growth media was then replaced with the same media containing TRAIL (Alexis Biochemicals), with or without 2.5 μg ml^{-1} cycloheximide (Sigma-Aldrich), and images were acquired at 3-min intervals for an additional 8 h. Cells still alive at the end of the 8 h were considered to have survived the treatment. MFC10A were imaged as described[11] but at ~5% CO_2 in phenol-red-free assay media[28] without EGF or insulin, to reduce cell migration. HeLa cells co-expressing IMS-RP and BID–GFP and H1299 FLIP–YFP cells were imaged[11] in the same media as described earlier for Fig. 1b–e but at ×20 magnification with frames every 3 or 10 min, respectively. H1666, SKBR3 and primary liver cells were imaged in 96-well glass bottom plates (Matrical) on a Nikon TE2000E at ×10 magnification in a 37 °C chamber with 5% CO_2.

Image analysis. Sister-cell tracking was performed manually. To assess whether there is a cell cycle effect on death time, we plotted death time as a function of time since division, as a proxy for cell cycle phase. Because the distribution of time since division is not uniform, one might infer a cell cycle effect but, in fact, a cell cycle effect would appear as a slope in the data, which we do not observe. For sister cell experiments in which CO_2-independent medium was used, there was more proliferation early in the sister-cell tracking movie (and thus more points near 'G2') because CO_2-independent medium does not support very long-term proliferation (division slows down, but the cells are not dying). Notably, the lack of a cell cycle effect is also apparent for sister cell experiments performed in DMEM—the standard medium for HeLa cells (Supplementary Fig. 5). In these experiments, the cells proliferated continuously during the sister-cell tracking movie. Cells that divided early in the movie often divided again if they survived stimulation with TRAIL, producing 'cousins' that were excluded from the analysis. This resulted in fewer cells in the G2 region, but this does not indicate a cell cycle effect, as the probability of surviving or dying is constant.

Analysis of IC-RP and EC-RP cleavage and IMS-RP translocation was performed as described[11]. To derive estimates for k_{IC} and θ, individual cell trajectories were fit with an equation derived from mathematical reduction of the differential equation model for the pathway (J.M.B., J.G.A, S.L.S., D. Lauffenburger and P.K.S., manuscript in preparation; see also Supplementary Fig. 5). The equations for the fitted relationship between T_d and k_{IC} and between T_d and θ were: $T_d = 6.9e^{(-53.7k_{IC})}$ and $T_d = 6.8\ \theta + 1.1$. FLIP–YFP fluorescence was quantified at $t = 0$ h (time of TRAIL addition) by manually outlining the cell and measuring the average fluorescence intensity within the outline. For FLIP–YFP cells, the time of death was scored as the first frame in which a cell exhibited apoptotic morphology. The slight pro-apoptotic effect observed with FLIP levels could be owing to its role as an activator of initiator caspases at low FLIP/CASP8 ratios[29]. BID–GFP fluorescence was quantified at $t = 0$ h by measuring the average fluorescence intensity from a representative area within the cell. To determine the absolute number of proteins per cell, the average BID–GFP fluorescence intensity from these movies was set equal to the average number of GFP-tagged proteins per cell as measured by quantitative immunoblot (Supplementary Fig. 9).

Flow cytometry. The distributions of initial protein levels were measured in HeLa cells (fixed with paraformaldehyde and permeabilized with methanol) on a FACSCalibur (BD Biosciences). The antibodies were carefully validated by knockout or knockdown and/or by overexpression of GFP-tagged fusion proteins. The following antibodies were used: anti-BID (HPA000722, Atlas Antibodies), anti-BAX (MAB4601, Chemicon International), anti-BCL2 (SC7382, Santa Cruz Biotechnology), anti-XIAP (610717, BD Biosciences) and anti-CASP3 (SC7272, Santa Cruz Biotechnology).The coefficient of variation of cells of similar size (as estimated by forward scatter) ranged from 0.21 to 0.28.

Modelling. The details of methods used will be described elsewhere (S.G., S.L.S., W. W. Chen and P.K.S., manuscript in preparation). In brief, a series of 10^4 simulations of the EARMv1.1 ordinary differential equation model[12] (modified to include general protein synthesis and degradation) were run in Jacobian (Numerica Technology). In previous work, only single values for each protein concentration were used[12]. Here, for each run of the model (representing one cell), initial protein levels were independently sampled from log-normal distributions having mean values as listed in Supplementary Table 1, and coefficients of variation as measured by flow cytometry (Fig. 2a) or set to 0.25 for proteins that were not measured. Initial protein concentrations and parameter values are listed in Supplementary Tables 1 and 2.

28. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**, 256–268 (2003).
29. Boatright, K. M., Deis, C., Denault, J. B., Sutherlin, D. P. & Salvesen, G. S. Activation of caspases-8 and -10 by $FLIP_L$. *Biochemical journal* **382**, 651–657 (2004).

LETTERS

Syk kinase signalling couples to the Nlrp3 inflammasome for anti-fungal host defence

Olaf Gross[1,2]*, Hendrik Poeck[1]*, Michael Bscheider[3], Catherine Dostert[2], Nicole Hannesschläger[1], Stefan Endres[3], Gunther Hartmann[4], Aubry Tardivel[2], Edina Schweighoffer[5], Victor Tybulewicz[5], Attila Mocsai[6], Jürg Tschopp[2] & Jürgen Ruland[1]

Fungal infections represent a serious threat, particularly in immunocompromised patients[1]. Interleukin-1β (IL-1β) is a key pro-inflammatory factor in innate antifungal immunity[2]. The mechanism by which the mammalian immune system regulates IL-1β production after fungal recognition is unclear. Two signals are generally required for IL-1β production: an NF-κB-dependent signal that induces the synthesis of pro-IL-1β (p35), and a second signal that triggers proteolytic pro-IL-1β processing to produce bioactive IL-1β (p17) via Caspase-1-containing multiprotein complexes called inflammasomes[3]. Here we demonstrate that the tyrosine kinase Syk, operating downstream of several immunoreceptor tyrosine-based activation motif (ITAM)-coupled fungal pattern recognition receptors, controls both pro-IL-1β synthesis and inflammasome activation after cell stimulation with *Candida albicans*. Whereas Syk signalling for pro-IL-1β synthesis selectively uses the Card9 pathway, inflammasome activation by the fungus involves reactive oxygen species production and potassium efflux. Genetic deletion or pharmalogical inhibition of Syk selectively abrogated inflammasome activation by *C. albicans* but not by inflammasome activators such as *Salmonella typhimurium* or the bacterial toxin nigericin. Nlrp3 (also known as NALP3) was identified as the critical NOD-like receptor family member that transduces the fungal recognition signal to the inflammasome adaptor Asc (Pycard) for Caspase-1 (Casp1) activation and pro-IL-1β processing. Consistent with an essential role for Nlrp3 inflammasomes in antifungal immunity, we show that Nlrp3-deficient mice are hypersusceptible to *Candida albicans* infection. Thus, our results demonstrate the molecular basis for IL-1β production after fungal infection and identify a crucial function for the Nlrp3 inflammasome in mammalian host defence *in vivo*.

IL-1β is a central orchestrator of immunity against various classes of pathogens, and a key trigger of inflammatory diseases. Its production is mediated by pro-Caspase-1-processing inflammasomes that contain danger sensors such as Nlrp proteins or Nlrc4 (also known as Ipaf) to connect upstream signals to Caspase-1 activation[3]. Although rapid progress has been made in identifying inflammasome components, the mechanisms upstream of inflammasome activation are not well understood.

Opportunistic fungi are clinically important pathogens[1] that cause life-threatening infections in immunocompromised individuals. However, immune responses to fungi are ill characterized, and whether and how fungi activate inflammasomes is unknown. We have co-incubated murine bone-marrow-derived dendritic cells (BMDCs) with *C. albicans* and measured the production of IL-1β. As a control, we activated inflammasomes with several well-characterized stimuli, including lipopolysaccharide (LPS) and ATP[4]. Live but not inactivated *C. albicans* induce a robust production of IL-1β (Fig. 1a and Supplementary Fig. 1a–d). Murine bone-marrow-derived macrophages (BMDMs), human peripheral blood monocytes (PBMCs) or THP-1 cells showed similar responses, excluding cell type or species-specific effects (Supplementary Fig. 1e–g and data not shown). Cell stimulation with the fungus directly activated Caspase-1, as detected by the appearance of the p10 cleavage product (Fig. 1d). Because Caspase-1-deficient BMDCs or BMDCs treated with the caspase inhibitor z-VAD-fmk have severe defects in IL-1β production after *C. albicans* stimulation, despite normal intracellular pro-IL-1β accumulation or secretion of the inflammasome-independent cytokine TNF-α (Fig. 1a–c and Supplementary Fig. 1h), we conclude that *C. albicans* induces an IL-1β-dependent inflammatory response mediated by Caspase-1.

Several ITAM-containing or ITAM-coupled C-type lectins, including dectin-1 (encoded by *Clec7a*), dectin-2 (*Clec4n*), Mincle (*Clec4e*) and potentially others, were recently identified as signalling fungal pattern recognition receptors that trigger pro-inflammatory cytokine responses[5,6]. Although there is redundancy at the receptor

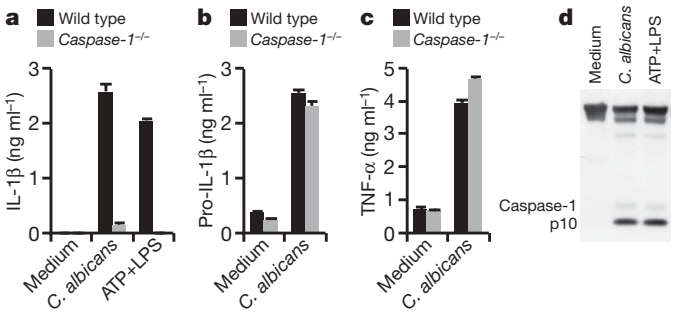


Figure 1 | *Candida albicans* activates a Caspase-1-dependent IL-1β response. a–c, BMDCs from wild-type or *Caspase-1*$^{-/-}$ mice were stimulated with *C. albicans* (5×10^6 cells ml^{-1}) for 5 h, or with ATP (5 mM) for 2 h after LPS prestimulation (0.5 ng ml^{-1} for 3 h). Secreted IL-1β (**a**), intracellular pro-IL-1β (**b**), or secreted TNF-α (**c**) were determined by ELISA. **d**, BMDCs were either left unstimulated (medium) or stimulated with *C. albicans* or with ATP after LPS priming. Caspase-1 activation was analysed by western blot, using an antibody against the Caspase-1 p10 cleavage product. Values in **a–c** are means and s.d. All results are representative of at least three independent experiments.

[1]III. Medizinische Klinik, Klinikum rechts der Isar, Technische Universität München, Ismaninger Str. 22, 81675 Munich, Germany. [2]Department of Biochemistry, University of Lausanne, Chemin des Boveresses 155, 1066 Epalinges, Switzerland. [3]Division of Clinical Pharmacology, Department of Internal Medicine, Ludwig-Maximilian Universität München, Germany. [4]Institute of Clinical Chemistry and Pharmacology, Universitätsklinikum Bonn, 53127 Bonn, Germany. [5]National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK. [6]Semmelweis University School of Medicine, Budapest, Hungary.
*These authors contributed equally to this work.

level, Syk and its downstream adaptor Card9 are both essential for the production of cytokines such as TNF-α in response to fungal recognition[7–9]. To test whether these signal transducers are required for IL-1β production, we analysed responses in Syk-deficient BMDCs[9]. To exclude Syk-dependent effects on cell differentiation, we also inhibited Syk kinase activity in wild-type BMDCs with the specific small molecule Syk inhibitor R406 that is in clinical trials for inflammatory diseases[10,11]. Syk deletion or inhibition blocked intracellular pro-IL-1β accumulation as well as TNF-α production in response to *C. albicans* (Fig. 2a, b and Supplementary Fig. 2a, b), indicating that Syk activity controls the first signal for IL-1β production after fungal recognition.

Notably, we also found that Syk blockade inhibited Caspase-1 activation by *C. albicans*, although the bacterial toxin nigericin induced normal Caspase-1 processing in $Syk^{-/-}$ cells (Fig. 2c, d). Moreover, Syk deletion or inhibition blocked *C.-albicans*-induced IL-1β secretion, and the priming of $Syk^{-/-}$ or R406-treated cells with LPS, to ensure Syk-independent pro-IL-1β production, did not rescue these defects (Fig. 2e–g). In contrast, Syk blockade does not affect IL-1β production after treatment with ATP or nigericin that both activate Nlrp3 inflammasomes, or in response to *S. typhimurium* that activates Nlrc4 inflammasomes[12], excluding a general function for Syk in Caspase-1 activation (Fig. 2e, f, h and Supplementary Fig. 2c). Together, these data indicate that Syk kinase signalling controls both pro-IL-1β production and inflammasome activation specifically after fungal recognition. Because inflammasome activity is also required for pro-IL-18 processing[3,13], we also tested the role of Syk signalling in IL-18 secretion. In line with the data described earlier, LPS-primed $Syk^{-/-}$ cells do not secrete measurable amounts of IL-18 after stimulation with *C. albicans*, although they produce normal IL-18 levels in response to ATP or nigericin (Fig. 2i).

We next tested the role of Card9 in these pathways. $Card9^{-/-}$ BMDCs[7] also failed to produce IL-1β after *C. albicans* stimulation (Fig. 2j). Consistent with the requirement of Card9 in Syk-mediated NF-κB activation[7,8] pro-IL-1β synthesis was blocked in $Card9^{-/-}$ cells (Fig. 2k). Yet, in contrast to Syk inhibition, Card9 deletion did not effect *C.-albicans*-triggered Caspase-1 activation (Fig. 2l), and after LPS priming $Card9^{-/-}$ cells secreted mature IL-1β into the culture supernatant (Fig. 2l). Thus, the Card9 pathway transduces the Syk signal selectively to pro-IL-1β synthesis, but Card9 is dispensable for inflammasome activation. Toll-like receptor (TLR)-activated MyD88 signalling is also not essential for inflammasome activation by *C. albicans* (Supplementary Fig. 3a, b).

To determine the mechanisms by which *C. albicans* might activate Caspase-1, we considered the possibility that the fungus may indirectly trigger inflammasomes by inducing the release of ATP or other activators from dying cells[14]. However, *C. albicans* did not cause substantial cell injury (Supplementary Fig. 4a). By stimulating BMDCs from mice lacking the ATP receptor $P2X_7$ (also known as $P2RX_7$)[4], we additionally excluded a critical requirement for cellular ATP release in IL-1β production by *C. albicans* (Fig. 3a).

Most inflammasome activators trigger cellular intermediary signals that couple to the activation of NLR proteins[3]. Common mechanisms implicated in inflammasome activation include K^+ efflux[15], lysosomal damage with release and activation of cathepsin B[16,17], and reactive oxygen species (ROS) production[15,18,19]. To study the role of these signalling cascades in *C. albicans* inflammasome activation, we first blocked potassium channels with glibenclamide[20]. *Candida-albicans*-dependent IL-1β production was inhibited by blocking K^+ efflux although the secretion of TNF-α was largely unaffected (Fig. 3b, c). Next, the lysosomal cathepsin B pathway was blocked by inhibiting lysosomal acidification with bafilomycin A[17]. In parallel, we inhibited the NADPH-oxidase-dependent ROS system with (2*R*,4*R*)-4-aminopyrrolidine-2,4-dicarboxylate (APDC)[18] (Fig. 3d, e). Neither of these inhibitors affected *C.-albicans*-dependent TNF-α secretion (Fig. 3d). Bafilomycin A also did not influence IL-1β production,

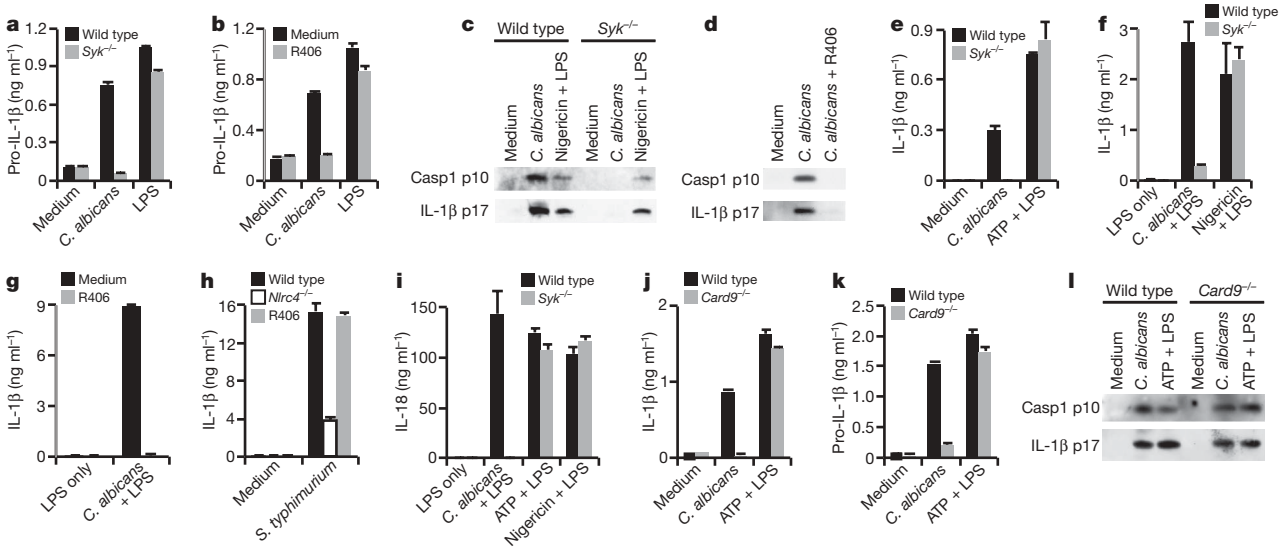


Figure 2 | Syk signalling controls pro-IL-1β synthesis and Caspase-1 activation after *C. albicans* stimulation. **a**, **b**, Wild-type or $Syk^{-/-}$ BMDCs, untreated or treated with 1 μM of the Syk inhibitor R406, were stimulated with *C. albicans* or LPS (100 ng ml^{-1}) before the measurement of intracellular pro-IL-1β. **c**, BMDCs from wild-type or $Syk^{-/-}$ mice were stimulated with *C. albicans* or with nigericin after LPS priming. Supernatants were analysed by western blot for the presence of Caspase-1 p10 and IL-1β p17 cleavage products. **d**, BMDCs were stimulated with *C. albicans*, with or without 1 μM R406 pretreatment. Processing of Caspase-1 and secretion of mature IL-1β p17 was analysed as in **c**. **e**, BMDCs from wild-type or $Syk^{-/-}$ mice were stimulated with *C. albicans* or with ATP and LPS before the measurement of IL-1β release. **f**, Wild-type or $Syk^{-/-}$ BMDCs were prestimulated with LPS and stimulated with *C. albicans* or nigericin before the measurement of IL-1β production. **g**, BMDCs were prestimulated with LPS, untreated or treated with R406 (30 min), and left unstimulated or stimulated with *C. albicans*. IL-1β production was measured by ELISA. **h**, BMDCs from wild-type mice, pretreated with R406 as indicated, or from $Nlrc4^{-/-}$ mice were stimulated with *S. typhimurium* (multiplicity of infection (m.o.i.) 10) for 5 h. IL-1β production was determined by ELISA. **i**, Cells were stimulated as in **e** and **f**, and IL-18 production was determined in the supernatants. **j**, **k**, BMDCs from wild-type or $Card9^{-/-}$ mice were stimulated with *C. albicans* or ATP and LPS. Secreted IL-1β (**j**) or intracellular pro-IL-1β (**k**) was determined as above. **l**, BMDCs from wild-type or $Card9^{-/-}$ mice were stimulated with *C. albicans* or ATP after LPS priming. Processing and secretion of Caspase-1 (p10) or IL-1β (p17) were determined by western blot. All values in **a**, **b** and **e**–**k** are means and s.d. Results are representative of at least three independent experiments.

suggesting that the lysosomal cathepsin B pathway is not involved in the *C. albicans* response (Fig. 3e). Consistently, BMDCs from cathepsin-B-deficient mice[21] also produce normal amounts of IL-1β after *C. albicans* exposure (Supplementary Fig. 4b). In contrast, ROS inhibition with APDC impaired *C.-albicans*-dependent IL-1β secretion in a dose-dependent manner (Fig. 3e and Supplementary Fig. 4c). Caspase-1 processing by *C. albicans* was also inhibited in cells that were treated with APDC, but not in those that received bafilomycin A (Fig. 3f and Supplementary Fig. 4d), suggesting that *C.-albicans*-dependent inflammasome triggering involves ROS production. In support of this, IL-1β secretion by *C. albicans* was impaired after NADPH oxidase inhibition with dipheylene iodoinium (Supplementary Fig. 4e)[18].

Previous studies with $Syk^{-/-}$ cells have demonstrated the essential role of Syk in ROS generation after fungal recognition[22]. Similarly, Syk inhibition with R406 reduced ROS production by *C. albicans* to a comparable degree to APDC treatment (Fig. 3g). In contrast, $Card9^{-/-}$ cells that show regular Caspase-1 activation also have normal ROS production after exposure to *C. albicans* (Supplementary Fig. 4g). Collectively, these results indicate that *C. albicans* bypasses the lysosomal cathepsin B pathway but triggers Syk-dependent ROS production and a K^+-efflux-dependent mechanism for inflammasome activation.

The molecular nature of the inflammasomes that are engaged by fungi is completely uncharacterized. Because ROS production and K^+ efflux have been linked to the activation of Nlrp3-inflammasome complexes in responses to diverse inflammatory triggers including ATP, asbestos and others[15,18,19], we proposed that *C. albicans* might engage Nlrp3. Indeed, Nlrp3-deficient cells failed to activate Caspase-1 or to produce IL-1β specifically after *C. albicans* or ATP treatment, but not after transfection with the Nlrp3-independent inflammasome activator double-stranded DNA (poly(dA:dT)[20]; Fig. 4a and Supplementary Fig. 5a, b). In line with the essential role for the adaptor Asc in coupling activated Nlrp3 or DNA recognition to Caspase-1, Asc-deficient cells also failed to process Caspase-1 or IL-1β in response to *C. albicans*, ATP or poly(dA:dT) (Fig. 4b and Supplementary Fig. 5c, d). However, in contrast to Nlrp3 and Asc, Nlrc4 (ref. 12) is dispensable for inflammasome activation by *C. albicans* (Fig. 4c). We also tested BMDCs from other Nlrp-deficient (Nlrp6 and Nlrp12; J.T., unpublished) or mutant

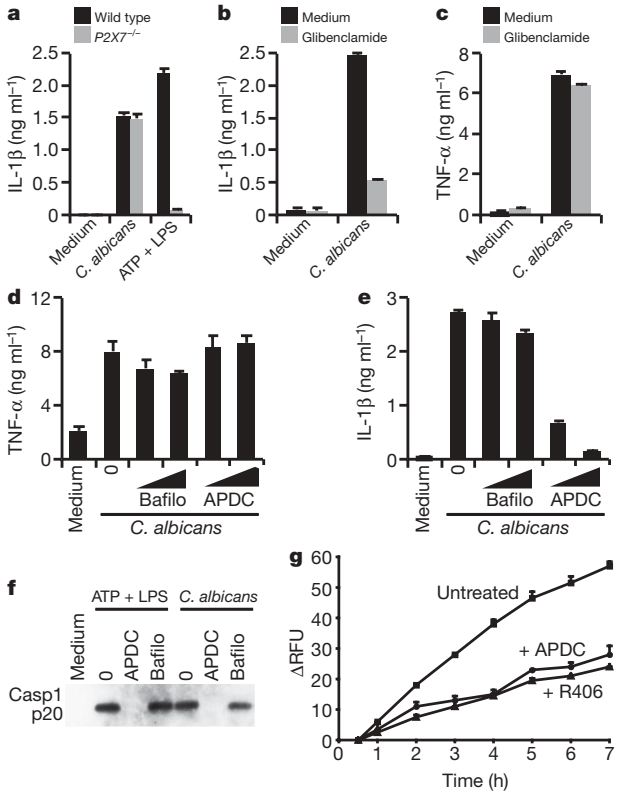


Figure 3 | Inflammasome activation by *C. albicans* involves ROS production and K^+ efflux. **a**, BMDCs from wild-type or $P2X7^{-/-}$ mice were stimulated with *C. albicans* or with ATP and LPS. IL-1β secretion was determined by ELISA. **b**, **c**, BMDCs were stimulated with *C. albicans* in the presence or absence of glibenclamide pretreatment. IL-1β (**b**) and TNF-α (**c**) secretion were determined by ELISA. **d**, **e**, BMDCs were left untreated (0) or treated with bafilomycin A (bafilo; 125 or 500 nM) or APDC (25 or 100 μM), and stimulated with *C. albicans*. TNF-α (**d**) and IL-1β (**e**) concentrations in the supernatant were determined by ELISA. **f**, BMDC were pretreated with bafilomycin A or APDC as in **c** and left untreated or stimulated with ATP and LPS or with *C. albicans*. Caspase-1 activation was determined by western blotting using an antibody that detects the processed p20 subunit. **g**, BMDCs were stimulated with *C. albicans* in the presence or absence of 15 min pretreatment with the ROS inhibitor APDC or the Syk inhibitor R406. ROS production was determined using the fluorescent probe 2′7′-dichlorofluorescin diacetate (H_2DCFDA). RFU, relative fluorescent units. All values in **a**–**e** and **g** are means and s.d. All results are representative of at least three independent experiments.

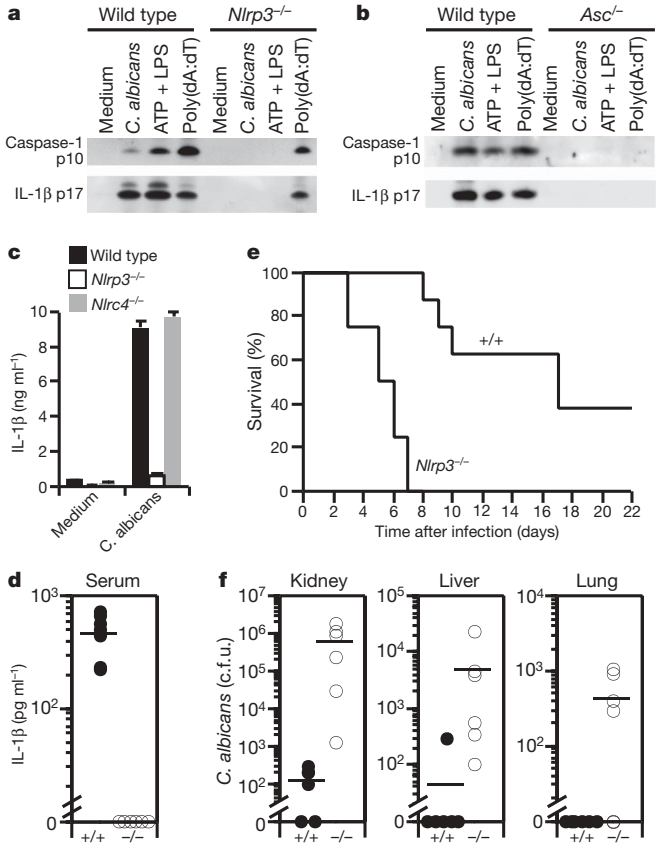


Figure 4 | The Nlrp3 inflammasome controls anti-fungal immunity. **a**, BMDCs from wild-type or $Nlrp3^{-/-}$ mice were stimulated with *C. albicans* or with ATP and LPS, or transfected with 2 μg ml^{-1} poly(dA:dT). Processing of Caspase-1 (p10) and production of IL-1β (p17) was determined by western blotting. **b**, BMDCs from wild-type or $Asc^{-/-}$ mice were treated and analysed as in **a**. **c**, BMDCs from wild-type, $Nlrp3^{-/-}$ or $Nlrc4^{-/-}$ mice were stimulated with *C. albicans*. IL-1β production was determined by ELISA. All values in **a**–**c** are means and s.d. All results are representative of at least three independent experiments. **d**–**f**, $Nlrp3^{-/-}$ and wild-type control mice were intravenously infected with *C. albicans* (10^5 colony forming units (c.f.u.)). **d**, IL-1β concentration in the serum of seven wild-type (+/+) and six $Nlrp3^{-/-}$ mice was determined after 4 h by ELISA. **e**, The frequency of viable mice is indicated over time. One out of two independent experiments with a total of 19 wild-type and nine $Nlrp3^{-/-}$ mice is shown. Statistical analysis was performed by log-rank test ($P < 0.0001$). **f**, Six wild-type and six $Nlrp3^{-/-}$ were killed 4 days after infection. *Candida albicans* titres were determined in the kidneys, livers and lungs.

mice (C57/BL6 that lack Nlrp1)[23] but found no defects in their capacity to respond to *C. albicans* (data not shown), indicating that *C. albicans* specifically engages the Nlrp3 inflammasome. To test whether Nlrp3 has a more general role in fungal responses, we also stimulated BMDCs with *Saccharomyces cerevisiae*. Similar to *C. albicans*, *S. cerevisiae* activates the inflammasome and induces IL-1β production in an Nlrp3-dependent manner (Supplementary Fig. 6a, b).

The pathophysiological role of Nlrp3 in inflammatory disorders is well established[3]. However, the *in vivo* function of Nlrp3 in host defence is largely unknown. We thus investigated the role of Nlrp3 *in vivo* by infecting mice with *C. albicans*. *Nlrp3*$^{-/-}$ mice have a severe defect in IL-1β production (Fig. 4d). Moreover, all Nlrp3-deficient animals died within 6 days after infection, whereas more than 50% of the control mice survived the challenge for over 16 days (Fig. 4e), consistent with the essential role for IL-1β in anti-fungal defence[2]. In separate experiments, we killed the animals 4 days after infection and assessed intravital fungal growth. Compared to wild-type mice, *Nlrp3*$^{-/-}$ mice had a more than 100–1,000-fold higher *C. albicans* load in the kidneys, livers and lungs (Fig. 4f), demonstrating that the Nlrp3 inflammasome mediates anti-fungal host defence *in vivo*.

In conclusion, we demonstrate that fungi can activate inflammasomes, and show, to our knowledge, the first essential role for Nlrp3 in host protection. Moreover, we describe a new mechanism of inflammasome activation that involves Syk kinase signalling. This finding may have broader implications. Syk-coupled C-type lectin receptors are emerging as important activators of inflammatory responses, and can detect exogenous or endogenous ligands[6,24,25]. In addition, prototypic pro-inflammatory crystals such as uric acid particles, which are responsible for Nlrp3-dependent gout[14], activate Syk by direct lipid membrane binding[26]. It will thus be important to study the functions of Syk in inflammasome activation in a broader context and to investigate whether the beneficial clinical effects of Syk inhibitors in inflammatory disorders[11,27] may, at least in part, be due to inhibition of the inflammasome.

METHODS SUMMARY

Mouse lines have been previously described[4,7,12,14,21,28,29]. Cells were stimulated in OptiMEM serum-free medium (Invitrogen) at 1×10^6 cells ml^{-1} in 12- or 6-well plates. Unless otherwise stated, cells were stimulated for 4 h with 5×10^6 live yeast cells per ml or with ATP (5 mM) or nigericin (3.4 μM) for 1 to 2 h. Where indicated, cells were primed with ultrapure LPS (*E. coli* K12, Invivogen, 0.5–5 ng ml^{-1}) for 3 h before addition of inflammasome activators[12]. Chemical inhibitors were added 30 min before inflammasome activation and 2.5 h after LPS priming, where applicable. Cell-free supernatants were analysed for cytokine secretion by ELISAs (BD, eBioscience or R&D) or subjected to western blot. Anti-mouse-IL-1β (R&D Systems), anti-mouse-Caspase-1 p10 (Santa Cruz, sc-514) and anti-mouse-Caspase-1 p20 (a gift from P. Vandenabeele) primary antibodies were used. For the determination of intracellular pro-IL-1β by ELISA, cells were lysed by repeated freeze–thaw cycles.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 February; accepted 12 March 2009.
Published online 1 April 2009.

1. Romani, L. Immunity to fungal infections. *Nature Rev. Immunol.* **4,** 1–23 (2004).
2. Vonk, A. G. *et al.* Endogenous interleukin (IL)-1α and IL-1β are crucial for host defense against disseminated candidiasis. *J. Infect. Dis.* **193,** 1419–1426 (2006).
3. Yu, H. B. & Finlay, B. B. The caspase-1 inflammasome: a pilot of innate immune responses. *Cell Host Microbe* **4,** 198–208 (2008).
4. Mariathasan, S. *et al.* Cryopyrin activates the inflammasome in response to toxins and ATP. *Nature* **440,** 228–232 (2006).
5. Bugarcic, A. *et al.* Human and mouse macrophage-inducible C-type lectin (Mincle) bind *Candida albicans*. *Glycobiology* **18,** 679–685 (2008).
6. Yamasaki, S. *et al.* Mincle is an ITAM-coupled activating receptor that senses damaged cells. *Nature Immunol.* **9,** 1179–1188 (2008).
7. Gross, O. *et al.* Card9 controls a non-TLR signalling pathway for innate anti-fungal immunity. *Nature* **442,** 651–656 (2006).
8. Hara, H. *et al.* The adaptor protein CARD9 is essential for the activation of myeloid cells through ITAM-associated and Toll-like receptors. *Nature Immunol.* **8,** 619–629 (2007).
9. LeibundGut-Landmann, S. *et al.* Syk- and CARD9-dependent coupling of innate immunity to the induction of T helper cells that produce interleukin 17. *Nature Immunol.* **8,** 630–638 (2007).
10. Braselmann, S. *et al.* R406, an orally available spleen tyrosine kinase inhibitor blocks Fc receptor signaling and reduces immune complex-mediated inflammation. *J. Pharmacol. Exp. Ther.* **319,** 998–1008 (2006).
11. Pine, P. R. *et al.* Inflammation and bone erosion are suppressed in models of rheumatoid arthritis following treatment with a novel Syk inhibitor. *Clin. Immunol.* **124,** 244–257 (2007).
12. Mariathasan, S. *et al.* Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf. *Nature* **430,** 213–218 (2004).
13. Mencacci, A. *et al.* Interleukin 18 restores defective Th1 immunity to *Candida albicans* in caspase 1-deficient mice. *Infect. Immun.* **68,** 5126–5131 (2000).
14. Martinon, F. *et al.* Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature* **440,** 237–241 (2006).
15. Petrilli, V. *et al.* Activation of the NALP3 inflammasome is triggered by low intracellular potassium concentration. *Cell Death Differ.* **14,** 1583–1589 (2007).
16. Halle, A. *et al.* The NALP3 inflammasome is involved in the innate immune response to amyloid-β. *Nature Immunol.* **9,** 857–865 (2008).
17. Hornung, V. *et al.* Silica crystals and aluminum salts activate the NALP3 inflammasome through phagosomal destabilization. *Nature Immunol.* **9,** 847–856 (2008).
18. Dostert, C. *et al.* Innate immune activation through Nalp3 inflammasome sensing of asbestos and silica. *Science* **320,** 674–677 (2008).
19. Cruz, C. M. *et al.* ATP activates a reactive oxygen species-dependent oxidative stress response and secretion of proinflammatory cytokines in macrophages. *J. Biol. Chem.* **282,** 2871–2879 (2007).
20. Muruve, D. A. *et al.* The inflammasome recognizes cytosolic microbial and host DNA and triggers an innate immune response. *Nature* **452,** 103–107 (2008).
21. Vasiljeva, O. *et al.* Reduced tumour cell proliferation and delayed development of high-grade mammary carcinomas in cathepsin B-deficient mice. *Oncogene* **27,** 4191–4199 (2008).
22. Underhill, D. M., Rossnagle, E., Lowell, C. A. & Simmons, R. M. Dectin-1 activates Syk tyrosine kinase in a dynamic subset of macrophages for reactive oxygen production. *Blood* **106,** 2543–2550 (2005).
23. Boyden, E. D. & Dietrich, W. F. Nalp1b controls mouse macrophage susceptibility to anthrax lethal toxin. *Nature Genet.* **38,** 240–244 (2006).
24. Robinson, M. J. *et al.* Myeloid C-type lectins in innate immunity. *Nature Immunol.* **7,** 1258–1265 (2006).
25. Chen, S. T. *et al.* CLEC5A is critical for dengue-virus-induced lethal disease. *Nature* **453,** 672–676 (2008).
26. Ng, G. *et al.* Receptor-independent, direct membrane binding leads to cell-surface lipid sorting and Syk kinase activation in dendritic cells. *Immunity* **29,** 807–818 (2008).
27. Weinblatt, M. E. *et al.* Treatment of rheumatoid arthritis with a Syk kinase inhibitor: a twelve-week, randomized, placebo-controlled trial. *Arthritis Rheum.* **58,** 3309–3318 (2008).
28. Li, P. *et al.* Mice deficient in IL-1 beta-converting enzyme are defective in production of mature IL-1 beta and resistant to endotoxic shock. *Cell* **80,** 401–411 (1995).
29. Kawai, T. *et al.* Unresponsiveness of MyD88-deficient mice to endotoxin. *Immunity* **11,** 115–122 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Peschel for helpful conversations, K. Schroder for critically reading the manuscript, RIGEL Inc. for providing the Syk inhibitor R406, P. Vandenabeele for the anti-Caspase-1 antibody and V. Dixit for *Nlrc4*$^{-/-}$ mice. O.G. is supported by a Marie Curie RTN ApopTrain Fellowship. This work was supported by Swiss National Science Foundation National Center of Competence in Research molecular oncology and Mugen grants to J.T., an EMBO long-term fellowship to C.D. and a Max-Eder-Program grant from Deutsche Krebshilfe and Sonderforschungsbereich grants from Deutsche Forschungsgemeinschaft to J.R.

Author Contributions O.G., H.P. and J.R. designed the research; O.G., H.P., M.B., N.H., C.D. and A.T. performed experiments; E.S., V.T. and A.M. contributed critical reagents, O.G., H.P., M.B., S.E., G.H., C.D., J.T. and J.R. analysed results; O.G. made the figures; O.G., H.P. and J.R. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.R. (jruland@lrz.tum.de).

METHODS

Mice. Mice deficient for *Caspase-1* (ref. 28), *Card9* (ref. 7), *P2X7* (ref. 4), *cathepsin B*[21], *Nlrp3* (ref. 14), *Asc*[12], *Nlrc4* (ref. 12) and *MyD88* (ref. 29) were used at 6–12 weeks of age and according to local guidelines. Syk-deficient embryonic liver chimaeras were generated as described[9].

Media and reagents. All reagents were supplied by Sigma unless otherwise stated. Cell culture reagents were from Invitrogen. FCS was from HyClone. Cytotox LDH-release assay kit was from Promega.

Cells. BMDCs, BMDMs and THP-1 cells were cultured as described[7]. Human PBMCs were isolated from whole blood of healthy, voluntary donors by standard Ficoll-Hypaque density gradient centrifugation (Biochrom).

Cell stimulation. Cells were stimulated in OptiMEM serum-free medium at 1×10^6 cells ml^{-1} in 12- or 6-well plates. *Candida albicans* cells were maintained on selective Chromagar plates (BD Biosciences). Before stimulation, they were expanded overnight on Columbia agar plates (BD Bioscience) at 30 °C. Viable *C. albicans* cells were rinsed off the plate, washed twice with PBS, left unopsonised and used directly for stimulation as described[7]. *Saccharamocyes cerevisiae* were prepared similarly. *Candida albicans* hyphae were prepared by incubating the fungal cells overnight in RPMI at 37 °C. Fungal yeast or hyphae were killed by incubation in a boiling water bath for 30 min. Unless otherwise stated, cells were stimulated for 4 h with 5×10^6 yeast cells per ml or equivalent amounts of hyphae. Cells were stimulated with ATP (5 mM) or nigericin (3.4 μM) for 1 to 2 h, unless otherwise stated. *Salmonella typhimurium* (m.o.i. 10)[12] and double-stranded DNA (poly(dA-dT)•poly(dA-dT), known as poly(dA:dT), 2 μg ml^{-1})[20] were used as described. Where indicated, cells were primed with ultrapure LPS (*E. coli* K12, Invivogen, 0.5–5 ng ml^{-1}) for 3 h before the addition of inflammasome activators as described[12]. Cell supernatants were analysed for cytokine secretion by ELISA (BD, eBioscience R&D Systems or MBL) or subjected to western blot analysis. For the determination of intracellular pro-IL-1β by ELISA, cells were lysed by repeated freeze–thaw cycles.

Chemical inhibitors. Chemical inhibitors were added 30 min before cell stimulation (2.5 h after LPS priming, where applicable). Pan-caspase-inhibitor (z-VAD-fmk, Calbiochem) was used at 25 ng ml^{-1}. R406, provided as a gift from Rigel, Inc. was used at 1 μM or as indicated. The ROS inhibitor APDC (Alexis) was used at 50 μM or as indicated. Bafilomycin A was used at 100 nM or as indicated. Dipheylene iodoinium was used at 10 μM, glibenclamide at 25 μM. As a control for specificity and toxicity, we confirmed that the inhibitors did not interfere with LPS- or *C.-albicans*-dependent TNF-α and/or pro-IL-1β production.

ROS assay. The ROS indicator H_2DCFDA (20 μM, Fluka) was added to the cells in HBSS 30 min before stimulation according to manufacturer's instructions. Fluorescence was recorded in 96-well plates over time with a Titertek FluoroskanII using a FITC filter (excitation 485 nm, emission 538 nm).

Western blotting. Cell supernatants or cell extracts were subjected to standard western blot techniques as described[20]. Proteins from cell-free supernatants were extracted by methanol/chloroform precipitation. Anti-mouse-IL-1β (R&D Systems), anti-mouse-Caspase-1 p10 (Santa Cruz, sc-514) and anti-mouse-Caspase-1 p20 (a gift from P. Vandenabeele) primary antibodies were used.

LETTERS

Precise genome modification in the crop species *Zea mays* using zinc-finger nucleases

Vipula K. Shukla[1], Yannick Doyon[2], Jeffrey C. Miller[2], Russell C. DeKelver[2], Erica A. Moehle[2], Sarah E. Worden[1], Jon C. Mitchell[1], Nicole L. Arnold[1], Sunita Gopalan[2], Xiangdong Meng[2], Vivian M. Choi[2], Jeremy M. Rock[2], Ying-Ying Wu[2], George E. Katibah[2], Gao Zhifang[1], David McCaskill[1], Matthew A. Simpson[1], Beth Blakeslee[1], Scott A. Greenwalt[1], Holly J. Butler[1], Sarah J. Hinkley[2], Lei Zhang[2], Edward J. Rebar[2], Philip D. Gregory[2] & Fyodor D. Urnov[2]

Agricultural biotechnology is limited by the inefficiencies of conventional random mutagenesis and transgenesis. Because targeted genome modification in plants has been intractable[1], plant trait engineering remains a laborious, time-consuming and unpredictable undertaking. Here we report a broadly applicable, versatile solution to this problem: the use of designed zinc-finger nucleases (ZFNs) that induce a double-stranded break at their target locus[2]. We describe the use of ZFNs to modify endogenous loci in plants of the crop species *Zea mays*. We show that simultaneous expression of ZFNs and delivery of a simple heterologous donor molecule leads to precise targeted addition of an herbicide-tolerance gene at the intended locus in a significant number of isolated events. ZFN-modified maize plants faithfully transmit these genetic changes to the next generation. Insertional disruption of one target locus, *IPK1*, results in both herbicide tolerance and the expected alteration of the inositol phosphate profile in developing seeds. ZFNs can be used in any plant species amenable to DNA delivery; our results therefore establish a new strategy for plant genetic manipulation in basic science and agricultural applications.

Current approaches to the challenge of improving agricultural productivity and global food production (for example, enhancing yield or engineering pest resistance) rely on conventional biotechnology approaches such as mutation breeding or transformation of novel genes into crop genomes. Both processes are inherently non-specific and relatively inefficient. Targeted genome modification in plant systems, which has been a long-standing but elusive goal, can overcome many of these logistical challenges. Previous efforts to drive targeted gene addition at endogenous loci in rice and *Arabidopsis* have relied on analysis of large numbers of transformants in order to recover extremely rare desired events[3].

Zinc finger nucleases (ZFNs) are a fusion of zinc-finger-based DNA recognition modules to an endonuclease domain[4]. ZFNs act by invoking the recombinogenic repair potential of a double-stranded break (DSB) in the DNA of living cells[5–7]. ZFN-induced DSBs enhance gene targeting at engineered loci in human cells[8] and in model plants[9,10]. Because the zinc-finger domain[11,12] can be engineered to recognize novel DNA sequences[13,14], ZFNs have been widely exploited for genome engineering at endogenous loci in eukaryotic systems (reviewed in ref. 2).

In this study, we assessed whether ZFN-driven gene addition could be used for trait engineering at an endogenous locus in maize. We targeted the *IPK1* gene, which encodes inositol-1,3,4,5,6-pentakisphosphate 2-kinase, an enzyme that catalyses the final step in phytate biosynthesis in maize seeds[15]. *IPK1* represents an attractive choice for targeting because phytate reduction is agriculturally important: phytate accounts for ~75% of total seed phosphorus[16], is an anti-nutritional component of feed grains and contributes to environmental pollution through the waste stream. Efforts to manipulate phytate accumulation via genetic modification have focused on reducing/eliminating the activities of enzymes that catalyse the conversion of inositol phosphate intermediates[17,18].

Two *Z. mays IPK* gene paralogues, here referred to as *IPK1* and *IPK2*, exist in the maize genome and share 98% sequence identity in the coding regions[15]; *IPK1* was selected for targeting based on its expression pattern. Using an archive of pre-validated 2-finger modules[19], we generated a panel of 66 ZFNs against 5 distinct positions of *IPK1*, focusing on the first two-thirds of the coding region and selecting sequences containing inter-paralogous single nucleotide polymorphisms (SNPs; Fig. 1a and Supplementary Fig. 1; see also Supplementary Table 6 for ZFN engineering and testing statistics). To overcome the low rates of DNA delivery to plant embryos or cultured cells[20,21], the ZFNs were initially assessed for efficacy using a mammalian reporter assay system[8] (Supplementary Fig. 2), followed by a yeast-based proxy system[22] (Fig. 1b). On the basis of this analysis, four ZFN pairs targeting the DNA sequences of Ile 71 and His 100 in exon 2 were selected for the specific editing of *IPK1* (Fig. 1a and Supplementary Table 1).

Repair of ZFN-induced DSBs by non-homologous end-joining (NHEJ) generates small deletions and insertions at the ZFN cleavage site[23,24], which provide a rapid indicator of ZFN activity at endogenous loci. In cultured maize cells that were transiently expressing ZFN 12 (but not controls), analyses of 6.5×10^4 chromatids revealed 28 deletions and 2 insertions aligning to the ZFN target site (Fig. 1c). This result illustrates that the human and yeast proxy system (Fig. 1b) identify ZFNs that induce a DSB at their intended target in plant cells.

Having established ZFN cleavage activity *in planta*, we used a selection-based scheme to disrupt *IPK1* by insertional gene addition. Two different donor constructs were generated, each containing short homology arms[19,25–27]: one carried an autonomous herbicide-tolerance gene expression cassette (*PAT*), whereas the second carried a non-autonomous donor that relied on precise trapping of the endogenous *IPK1* promoter for expression of the marker (Fig. 2a).

Four ZFN pairs designed to cleave *IPK1* at two positions in exon 2 were independently delivered to maize cells, along with either autonomous or non-autonomous donor plasmids. Transformed, herbicide-tolerant calli were genotyped at the *IPK1* locus (Fig. 2b and Table 1); our

[1]Dow AgroSciences, 9330 Zionsville Road, Indianapolis, Indiana 46268, USA. [2]Sangamo BioSciences, Point Richmond Tech Center, 501 Canal Boulevard, Suite A100, Richmond, California 94804, USA.

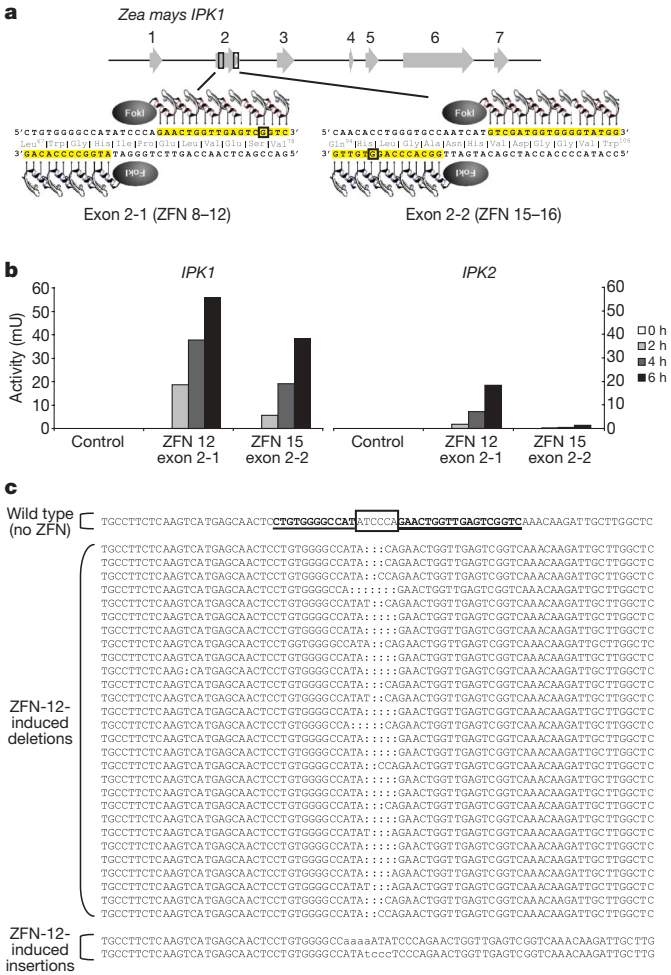


Figure 1 | ZFNs designed to target *Z. mays IPK1* induce DSBs at the endogenous locus. a, Schematic representation of the *Z. mays IPK1* gene. Exons are indicated by arrows; regions used as ZFN design templates are boxed. The magnified views illustrate binding sites for ZFN pairs used here; *IPK1*-specific SNPs are boxed (see Supplementary Fig. 1 for further details). **b**, Normalized ZFN reporter gene correction activity in a yeast proxy system[22]. Left: activity of indicated ZFNs against the maize *IPK1* locus. Right: activity of the same ZFNs against the *IPK2* locus. **c**, Assaying for ZFN-induced deletions at *IPK1*. Representative deep sequencing results of maize genomic PCR products from wild type (top line) or cells transiently expressing ZFN 12 are shown. Multiple deletions (colons) and insertions (lowercase) induced by NHEJ are aligned to the predicted cleavage site (box) delimited by the ZFN binding sites (underlined). Note that the chromatid carrying an insertion of four adenines could potentially represent a pyrosequencing artefact.

results revealed both monoallelic insertions of *PAT* at *IPK1* and events containing only a transgenic chromatid (Fig. 2b, lanes labelled 'TI/−' and 'TI/TI', respectively). Cloning and sequencing of genomic PCR products confirmed that gene addition had occurred in a precise, homology-directed manner (Supplementary Fig. 3). Analyses of all (~600) isolated herbicide-tolerant callus events are summarized in Table 1. All four ZFN pairs drove targeted gene addition into their target loci, albeit with different efficiencies. As expected, the non-autonomous donor strategy yielded fewer herbicide-tolerant events but significantly enriched this population for targeted versus random integration (Table 1).

Southern blot analysis of a representative panel of herbicide-tolerant events using an *IPK*-specific probe (Fig. 2c) is shown in Fig. 2d, top panel. In control and non-targeted herbicide-tolerant events (for example, 158, 413, 414, 416, 420), only the band corresponding to the non-disrupted allele of *IPK1* (>3.2 kb) was detected. In contrast, events carrying targeted integration of the *PAT* gene based on PCR genotyping (Fig. 2b) revealed one transgenic and one wild type, or a transgenic chromatid only. Of note, Southern blot analysis using an insert-specific probe (Fig. 2c) revealed multiple random integrations in non-targeted events (Fig. 2d, middle panel), but only a single band in a subset of correctly targeted events. On the basis of these results, we conclude that in these representative primary transformants, only directed transgene insertion occurred, with no random integration of the donor DNA. Finally, in leaf tissue of plants regenerated from correctly targeted calli, no randomly integrated ZFN expression construct was detected (Fig. 2d, bottom panel). Taken together, these data illustrate that ZFNs can drive targeted gene addition to an investigator-specified locus with no additional random integration of either donor DNA or the ZFN expression cassette.

In agreement with these genotyping data, *IPK1* transcript levels in primary callus tissue from selected targeted integration events were significantly reduced when compared to wild type (HiII) or random integration events (number 420; Fig. 3a). Callus number 419, which lacks a wild-type *IPK1* allele, accumulated the lowest levels of *IPK1* messenger RNA; residual signal may derive from wild-type cells in the sample because primary callus tissue transformed with silicon carbide whiskers remains chimaeric even after maintenance under selection.

We investigated whether ZFN-driven gene addition is associated with genome modifications at sequences other than the intended target. The off-target gene paralogue *IPK2* represents the closest sequence match in the maize genome for the binding site of each ZFN pair used here. We genotyped *IPK2* in T_0 plants carrying a transgene at *IPK1* (that is, events produced by functional ZFN expression). In five independent cases, we found the *IPK2* sequence to be wild type. These observations, along with yeast proxy data showing paralogue-specific action by the ZFNs (ZFN 15; Fig. 1b) and ultra-deep sequencing results that detected only wild-type *IPK2* chromatids in cell suspensions transiently expressing ZFNs (data not shown), suggest that the ZFNs show specificity for their intended paralogue. Formally, we cannot exclude the possibility that the ZFNs cleaved *in vivo* at *IPK2* but that the break was resolved via error-free NHEJ. In addition, we applied a SELEX/genotyping approach (Supplementary Fig. 4), which has been used to identify successfully rare but bona fide off-target sites for ZFNs that cleave the human *CCR5* gene[28]. Analysis of the five highest-probability off-target ZFN binding sites in the genomes of confirmed targeted integration events at *IPK1* revealed only wild-type sequences (Supplementary Tables 2–4).

We regenerated multiple independent, fertile plant lines from each of several genome-edited callus events. These plants were either self-pollinated or out-crossed to an inbred maize variety (DAS5XH751) lacking herbicide-tolerance genes. Individual progeny (T_1/F_1 plants generated by selfing/out-crossing, respectively) were genotyped at the *IPK1* locus (Fig. 3b). Random sampling of 9 T_1 seedlings from line 418-8 displayed a 2:6:1 (wild type:heterozygotic:homozygotic) segregation pattern for disruption of *IPK*, whereas analysis of 10 F_1 seedlings from line 418-6 revealed a 1:1 segregation of wild-type versus heterozygotic disruption at *IPK1* (Fig. 3b). This representative example shows that the ZFN-mediated genome modification observed in primary transformed callus and regenerated plants was transmitted to the next generation via normal sexual reproduction.

Consistent with these data, a proportion of the T_1/F_1 seedlings displayed an herbicide-tolerant phenotype: F_1 progeny of line 418-3 (monoallelic) yielded a 5:4 ratio of tolerant to susceptible plants, whereas T_1 progeny of event number 273 (monoallelic) yielded 25 herbicide-tolerant and 16 susceptible plants. Of note, all progeny (33 F_1 and 5 T_1 plants, derived from out-crossing and selfing, respectively) of lines derived from event number 419 survived selection, confirming this event as a biallelic targeted insertion.

We assessed the impact of ZFN-mediated insertional mutagenesis at *IPK1* on the accumulation of phytate and its precursors in individual seeds of genome-edited plants. *IPK1* transcript analysis in T_2

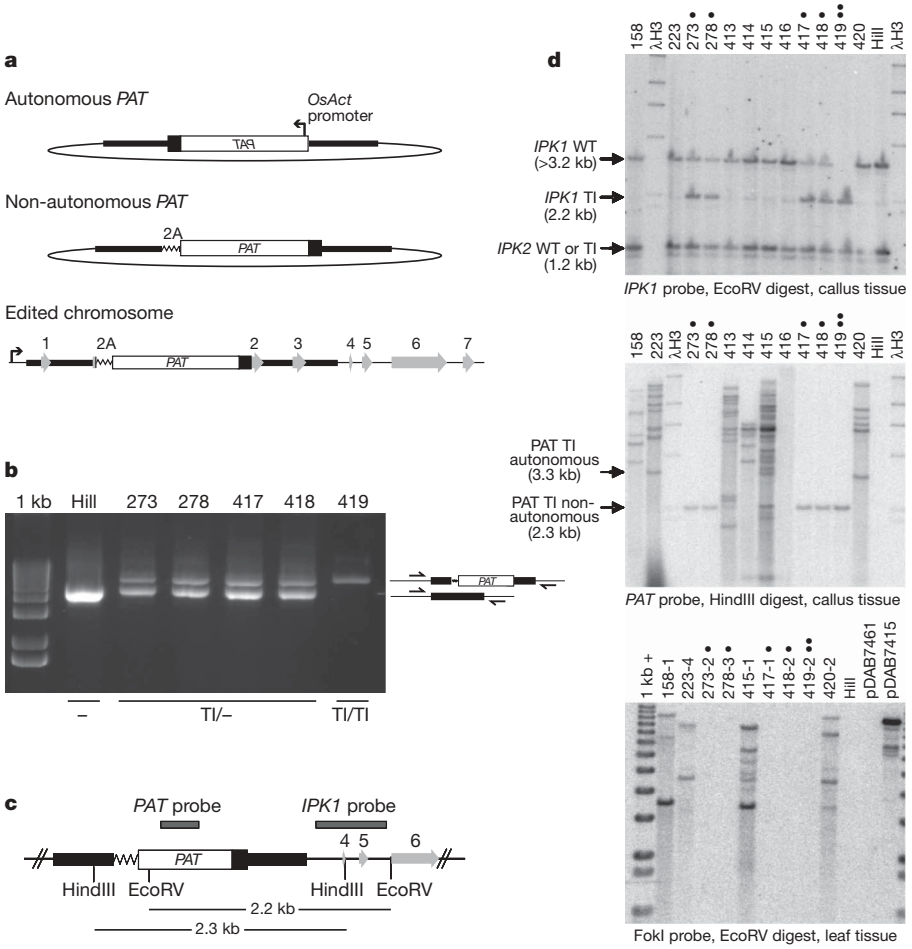


Figure 2 | Targeted gene addition to *IPK1* in maize. **a**, The top and middle panels are schematic representations of *Z. mays IPK1* donor plasmid constructs comprised of *PAT* gene cassettes flanked by short (815 bp) segments of *IPK1* sequence identity (thick black lines). *OsAct*, rice actin promoter; 2A, ribosomal stuttering signal. The bottom panel is an illustration of the expected edited chromosome containing the non-autonomous donor integrated into the endogenous *IPK1* locus. The bent arrow indicates the endogenous *IPK1* promoter. **b**, Representative results of PCR-based genotyping of the *IPK1* locus in maize calli. Right: positions of primers hybridizing to *IPK1* gene sequences outside of the donor homology region. From left to right, the labels at the top of the panel indicate molecular mass markers, wild-type HiII and targeted integration (TI) event numbers. At the bottom of the panel, TI/– indicates mono-allelic; TI/TI indicates bi-allelic insertions. **c**, Experimental schematic for Southern blot analyses of selected herbicide-tolerant events. Digestion scheme and predicted band sizes for hybridization at the *IPK1* locus are shown. Relative positions of the *IPK1* probe (539 nucleotides) and *PAT* gene probe (555 nucleotides) are indicated by bars. **d**, Southern blot analyses of genomic DNA (gDNA) from maize events. For all panels, molecular mass markers (λH3 or 1 kb+), wild-type (HiII) controls and event numbers are indicated in lane headers. Events designated as targeted integration based on PCR genotyping are indicated with dots. Probe, restriction enzyme and tissue type are indicated at the bottom of the panel. Top: interrogation of *IPK1* locus size. Hybridization bands corresponding to wild-type *IPK1* (*IPK1* WT; >3.2 kb), disrupted *IPK1* (*IPK1* TI; 2.2 kb) and the paralogous *IPK2* gene (1.2 kb) are indicated by arrows to the left. Middle: interrogation of *PAT* transgene copy number and integration site. Hybridization bands corresponding to targeted integration of *PAT* in either the autonomous (3.3 kb) or non-autonomous (2.3 kb) configuration at *IPK1* are indicated by arrows to the left. Bottom: interrogation of ZFN transgene copy number. Plasmid control lanes: negative, pDAB7461; positive, pDAB7415.

seeds from lines of event 418 confirmed that the effect of gene disruption on *IPK1* expression was transmitted through at least two rounds of meiosis (Fig. 3c). Next, we assayed seeds from ZFN-modified, control transgenic (that is, random integration of *PAT*) and wild-type maize plants for the relative ratios of multiple inositol phosphate species. As shown in Fig. 3d, segregating T_2 seeds from transgenic plants carrying a randomly inserted *PAT* gene show the expected (indistinguishable from wild type) distribution of high phytate (InsP6) and low inorganic phosphate (P_i) accumulation (Fig. 3d left and data not shown). In contrast, the relative distributions of InsP6 and P_i ratios in targeted integration event 418 show that a significant number of seeds in this segregating population display reduced phytate levels with a concomitant increase in inorganic phosphate (Fig. 3d, right). Similar observations were made in lines from each unique event analysed (data not shown).

Taken together, our data validate the hypothesis that *Z. mays IPK1* is a suitable gene target for manipulation of phytate accumulation and provide the basis for development of a reduced-phytate trait with agronomic and ecological significance. These results also show that disruption of *IPK1* via targeted gene addition using an herbicide-tolerant-encoding donor DNA couples phytate reduction with herbicide tolerance, thereby delivering a dual phenotype, or stacked trait, through targeted manipulation of a single locus. Moreover, we have failed to detect off-target ZFN-induced changes to the genome of T_0 plants. However, if such changes do occur, they can be rapidly eliminated through introgression of the desired allele into a different genetic background, a routine agricultural practice.

Our data describe targeted genome modification in *Zea mays*, an important crop that exemplifies all the genomic complexities typical of cereal grains[29]. An integrated experimental and bioinformatic analysis of current zinc-finger protein (ZFP) archives (Supplementary Information) indicates that this approach can be applied to any gene locus ≥1 kb in the maize genome. In fact, in a separate set of experiments, we have applied the approach used for *IPK1* to an unrelated

Table 1 | Frequency of targeted integration of *PAT* into *Z. mays IPK1*

Donor	ZFN number	ZFN target position	Total events*	Number of targeted integration events	Targeted integration (%)
Autonomous *PAT*	8	Exon 2-1	29	1	3.4
Autonomous *PAT*	12	Exon 2-1	31	4	12.9
Autonomous *PAT*	15	Exon 2-2	195	43	22.1
Autonomous *PAT*	16	Exon 2-2	216	46	21.3
Autonomous *PAT* + ZFN†	12	Exon 2-1	39	6	15.4
Autonomous PAT	N/A	N/A	25	0	0.0
Total			535	100	
Non-autonomous PAT	8	Exon 2-1	1	1	100.0
Non-autonomous PAT	12	Exon 2-1	5	3	60.0
Non-autonomous PAT	15	Exon 2-2	16	11	68.8
Non-autonomous PAT	16	Exon 2-2	8	4	50.0
Non-autonomous *PAT* + ZFN†	12	Exon 2-1	30	5	16.7
Total			60	24	

* Total number of herbicide-tolerant calli.
† Donor and ZFN present on the same plasmid.

and cytogenetically distinct locus in the maize genome, and found the efficiency of targeting to be comparable to that observed at *IPK1* (Supplementary Fig. 5 and Supplementary Table 5; see also Supplementary Table 6 for ZFN engineering and testing statistics).

The feasibility of extending this approach for genome modification to diverse plant species is supported by examples describing ZFN-mediated reporter-gene modification in tobacco and *Arabidopsis*[9,24] and editing of endogenous loci in tobacco[10,30]. More broadly, the use of direct-DNA delivery methods enables application of the approach to any plant species that is amenable to tissue culture and regeneration (including crops, vegetables and ornamentals), because the sole prerequisites for ZFN design are DNA sequence information and genome annotation. The combination of high-fidelity DNA recognition/cleavage by engineered ZFNs plus homology-directed repair at the specified break sites (a widely conserved biological pathway) makes precise modification of native genomes in plants practical and feasible for the first time. This approach, combined with rapid advances in genome sequencing technologies and bioinformatics, as well as the ongoing development of novel DNA delivery methods, establishes an efficient and precise strategy for plant genome engineering.

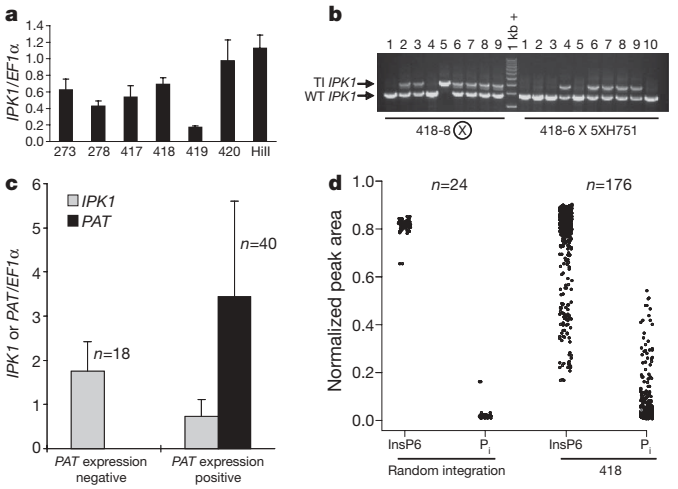


Figure 3 | ZFN-mediated gene disruption of *Z. mays IPK1* is stable and heritable. **a**, Quantification of *IPK1* mRNA in transformed callus tissue. *X*-axis labels indicate transgenic event number; *y*-axis indicates levels of *IPK1* mRNA normalized to that of *EF1α* mRNA as described in the Supplementary Information. Each bar represents 4–6 independent measurements; error bars indicate standard deviations of replicate measurements of the same mRNA sample. **b**, PCR genotyping of T_1/F_1 plants from event 418. Left, progeny from self pollination of line 418-8; right, progeny from out-crossing of line 418-6. Bands corresponding to wild type (WT) or targeted integration (TI) alleles of *IPK1* are indicated by arrows. **c**, Quantification of *IPK1* and *PAT* mRNA in T_2/F_2 seeds of event 418. *X*-axis labels indicate groupings of samples according to detection of *PAT* mRNA; *y*-axis indicates relative expression of *IPK1* or *PAT* mRNA normalized to *EF1α* mRNA as described in the Supplementary Information. *n*, number of individual seeds in each grouping; each seed was subjected to 4–6 independent measurements. Error bars indicate standard deviations of replicate measurements of the same mRNA sample. **d**, Normalized ratios of phytate (InsP6) and P_i in T_2/F_2 seeds of event 418. Data points indicate the distribution of InsP6 and P_i ratios (normalized peak area, see Supplementary Information) in multiple (*n*) seeds from a randomly inserted transgenic *PAT* line (left) and targeted integration event 418 (right). Each data point represents one of three independent measurements taken per seed.

METHODS SUMMARY

Zinc-finger nucleases directed against *Zea mays* genomic sequences were designed and validated as described in the Methods; further information regarding the full amino acid sequences of the ZFP portion of each ZFN used in this study and discussion of the design rationales are included in the Supplementary Information. The construction of ZFN and donor plasmids was carried out using standard molecular biology methods. Full descriptions of nucleic acid analyses including qRT–PCR and southern blotting, as well as descriptions of plant cell culture conditions, DNA delivery methods and phosphate/phytate analysis are detailed in the Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 October 2008; accepted 17 March 2009.
Published online 29 April 2009.

1. Puchta, H. Gene replacement by homologous recombination in plants. *Plant Mol. Biol.* **48**, 173–182 (2002).
2. Carroll, D. Progress and prospects: zinc-finger nucleases as gene therapy agents. *Gene Ther.* **15**, 1463–1468 (2008).
3. Iida, S. & Terada, R. Modification of endogenous natural genes by gene targeting in rice and other higher plants. *Plant Mol. Biol.* **59**, 205–219 (2005).
4. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl Acad. Sci. USA* **93**, 1156–1160 (1996).
5. Puchta, H. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.* **56**, 1–14 (2005).
6. Jasin, M. Genetic manipulation of genomes with rare-cutting endonucleases. *Trends Genet.* **12**, 224–228 (1996).
7. Bibikova, M. *et al.* Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol. Cell. Biol.* **21**, 289–297 (2001).
8. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).
9. Wright, D. A. *et al.* High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J.* **44**, 693–705 (2005).
10. Cai, C. Q. *et al.* Targeted transgene integration in plant cells using designed zinc finger nucleases. *Plant Mol. Biol.* **69**, 699–709 (2008).
11. Miller, J., McLachlan, A. D. & Klug, A. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J.* **4**, 1609–1614 (1985).
12. Pavletich, N. P. & Pabo, C. O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809–817 (1991).
13. Isalan, M. & Choo, Y. Rapid, high-throughput engineering of sequence-specific zinc finger DNA-binding proteins. *Methods Enzymol.* **340**, 593–609 (2001).
14. Pabo, C. O., Peisach, E. & Grant, R. A. Design and selection of novel Cys2his2 zinc finger proteins. *Annu. Rev. Biochem.* **70**, 313–340 (2001).
15. Sun, Y. *et al.* Inositol 1,3,4,5,6-pentakisphosphate 2-kinase from maize: molecular and biochemical characterization. *Plant Physiol.* **144**, 1278–1291 (2007).

16. Raboy, V. Seeds for a better future: 'low phytate' grains help to overcome malnutrition and reduce pollution. *Trends Plant Sci.* **6,** 458–462 (2001).
17. Stevenson-Paulik, J. *et al.* Generation of phytate-free seeds in *Arabidopsis* through disruption of inositol polyphosphate kinases. *Proc. Natl Acad. Sci. USA* **102,** 12612–12617 (2005).
18. Raboy, V. myo-Inositol-1,2,3,4,5,6-hexakisphosphate. *Phytochemistry* **64,** 1033–1043 (2003).
19. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).
20. Gordon-Kamm, W. J. *et al.* Transformation of maize cells and regeneration of fertile transgenic plants. *Plant Cell* **2,** 603–618 (1990).
21. Frame, B. R. *et al. Agrobacterium tumefaciens*-mediated transformation of maize embryos using a standard binary vector system. *Plant Physiol.* **129,** 13–22 (2002).
22. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature Biotechnol.* **26,** 702–708 (2008).
23. Bibikova, M. *et al.* Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. *Genetics* **161,** 1169–1175 (2002).
24. Lloyd, A. *et al.* Targeted mutagenesis using zinc-finger nucleases in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **102,** 2232–2237 (2005).
25. Porteus, M. H. Mammalian gene targeting with designed zinc finger nucleases. *Mol. Ther.* **13,** 438–446 (2006).
26. Moehle, E. A. *et al.* Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc. Natl Acad. Sci. USA* **104,** 3055–3060 (2007).
27. Lombardo, A. *et al.* Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature Biotechnol.* **25,** 1298–1306 (2007).
28. Perez, E. E. *et al.* Establishment of HIV-1 resistance in $CD4^+$ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnol.* **26,** 808–816 (2008).
29. Wei, F. *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.* **3,** e123 (2007).
30. Maeder, M. L. *et al.* Rapid ''open-source'' engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell* **31,** 294–301 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors wish to thank numerous colleagues at Dow AgroSciences and Sangamo BioSciences for discussions regarding this project, the DAS Gamma group for statistics support, Sangamo's production group for technical support and A. Klug for a critical reading of the manuscript. We are also grateful to W. Kleschick, D. Kittle and E. Lanphier for encouragement.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to V.K.S. (vkshukla@dow.com).

METHODS

ZFN design and proxy system validation. ZFNs directed against *Zea mays IPK1* and *Zp15* (Supplementary Information) were designed using an archive of pre-validated two-finger modules as described[31,32]. Design efforts for *IPK1* were focused on regions in exons 1–3 containing inter-paralogue SNPs (Supplementary Fig. 1). The full amino acid sequence of the ZFP portion of each ZFN used in this study is provided in Supplementary Tables 1 and 5; ZFN engineering and testing statistics are provided in Supplementary Table 6. The assembled ZFPs were tested for affinity and specificity of DNA binding using an ELISA assay[32]. ZFNs were tested for chromosomal reporter gene correction in budding yeast using an *in vivo* MEL1 reconstitution assay as described after induction of ZFN synthesis for 2, 4 and 6 h[32].

Plasmid construction. A stepwise modular cloning scheme was devised to assemble expression vectors for any given pair of ZFN-encoding genes. A vector including redesigned and synthesized segments of an NLS derived from maize *op-2* (RKRKESNRESARRSRYRK)[33] and a FokI nuclease domain using the maize codon bias was slightly modified to create an extra SacI site. A similar vector was modified to include the 2A sequence from the *Thosea asigna* virus (EGRGSLLTCGDVEENPGP)[34,35]. Cassettes encoding open reading frames of individual zinc-finger proteins were cloned into either of these vectors via KpnI and BamHI restriction sites, and the two vectors were then combined via BglII/XhoI restriction sites, yielding an intermediate construct including 2 ZFN-encoding domains flanked by NcoI and SacI restriction sites. The NcoI/SacI cassette was excised via restriction and ligated into the plasmid backbone pDAB3872, which contains a promoter from the maize ubiquitin-1 gene[36] and terminator sequences from maize root preferential cationic peroxidase gene (*Per5*, US patent 7,179,902). The resulting plasmids include the ZFN genes, selectable markers and flanking attL sites for convenient manipulation using the Gateway system from Invitrogen. Each of the ZFN constructs generated using this cloning scheme were transformed into *Escherichia coli* DH5α cells and subsequently maintained under the appropriate selection.

Donor constructs containing regions of homology to *Z. mays IPK1* were generated as follows: oligonucleotides specific to regions adjacent to the predicted ZFN cleavage sites (5′ region: 5′-GCGGCCGCGTCTCACCGCGGCTTGGGGATTGGATACGGAGCT-3′; 5′-ACTAGTGATATGGCCCCACAGGAGTTGCTCATGACTTG-3′; 3′ region: 5-ACTAGTCCAGAACTGGTTGAGTCGGTCAAACAAGATTGCT-3′; 5′-GTCGACCTTGATGCTACCCATTGGGCTGTTGT-3′) were used to amplify gDNA from maize variety HiII. Purified fragments were cloned into pCR2.1 plasmid, subsequently digested with NotI and SpeI (5′ region) or SpeI and SalI (3′ region) and ligated into base plasmid pBC SK(-). Cassettes comprising the herbicide-tolerance genes were inserted into the base vectors via SpeI digestion and ligation. In the case of the autonomous herbicide-tolerance gene donor, the promoter sequence is derived from *O. sativa* actin 1[37] (GenBank accessions S44221 and X63830). The herbicide-tolerance gene consists of a modified version of the *PAT* (phosphinothricin acetyl transferase) coding region originally derived from *Streptomyces viridochromogenes*[38] (GenBank accession I43995). The terminator sequences are derived from *Z. mays* L. lipase (GenBank accession L35913). In the case of the non-autonomous herbicide-tolerance gene donors, the promoter fragment was replaced with the 2A sequence described above.

Donor constructs containing regions of homology to *Zp15* (Supplementary Fig. 5) were generated as follows: oligonucleotides specific to regions adjacent to the predicted ZFN cleavage sites (5′ region: 5′-GCGGCCGCATGCAAGAGCTGTTGATC-3′; 5′-CAATTGCCGGCGTAGTAGGGCGCCGCCGCCAGC-3′; 3′ region: 5-CAATTGGTGTGGGCAGCCGAGCGCCATGTTCCAG-3′; 5′-GTCGACCGATACTGATGCGGACCGTCCACCTTGTC-3′) were used to amplify gDNA from maize variety HiII. Purified fragments were cloned into pCR2.1 plasmid, digested with MfeI and NotI (5′ region) or SpeI and SalI (3′ region) and ligated into base plasmid pBC SK(-). Cassettes comprising the autonomous herbicide-tolerance genes were subsequently inserted into the base vectors via MfeI digestion and ligation. In the case of the autonomous herbicide-tolerance gene donor, the promoter sequence is derived from *O. sativa* actin 1[37] (GenBank accessions S44221 and X63830). A cassette comprising the herbicide-tolerance gene was subsequently inserted into the base vectors via MfeI digestion and ligation. The promoter sequence was identical to that used for *IPK1*. The herbicide-tolerance gene consists of the *AAD* (aryloxyalkanoate dioxygenase) gene, which confers resistance to aryloxyphenoxypropionate herbicides (US patent PCT 2005/014737), modified to include a codon utilization pattern designed to optimize expression in plants. The terminator sequence was identical to that used for *IPK1*.

Nucleic acid analyses. All plasmids were transformed into and maintained in *E. coli* strains (One Shot Top 10 or MAX Efficiency DH5α) from Invitrogen Life Technologies, as per the manufacturer's recommendations. Maize genomic DNA isolation and all plasmid isolations were carried out using Qiagen kits as per the manufacturer's recommendations. Both ZFN-encoding and donor DNAs were directly delivered, as nonlinearized plasmids, into embryogenic cell cultures of maize via Whiskers (see below), followed by bulk gDNA isolation and PCR amplification of the target region 48 h after delivery. Deep sequencing of purified amplification products was carried out by 454 Life Sciences. All amplifications, sequencing reactions, Southern blotting, DNA digestions and agarose gel electrophoresis were carried out according to standard protocols[39].

Plant total RNA was isolated from 2–3-week-old cell cultures or 3-week-old seedlings using the Qiagen RNeasy kit according to the manufacturer's instructions and frozen in liquid N_2. Before column elution, DNA was removed by on-column DNase digestion. A portion of maize *EF1α* cDNA (*EF1α*) was cloned from maize genotype HiII by RT–PCR using forward primer 5′-GCATAGCCGTTGCCAATC-3′ and reverse primer 5′-CTCCCAGGCTGACTGTGCTG-3′ for use as internal standard for *PAT* and *IPK1* gene expression studies. Primers and probes for Taqman qRT–PCR are as follows: EF1α (forward primer 5′-TCCTTCACAATCTCTTCATAACGTG-3′; reverse primer 5′-TGGTTTTGAGGCTGGTATCTCC-3′; probe (5′ FAM, 3′ TAMRA) 5′-CTGCTGCAACAAGATGGATGCAACCACTCCCAAATAC-3′); *IPK1* (forward primer 5′-TGTTCTTAGCAGCCGTCCTGCTG-3′; reverse primer 5′-GGTGCTGATGCATCTTATATCTCGTTACTAG-3′; probe (5′ FAM, 3′ TAMRA) 5′-CCACTCTTTATTTTCTGGCAATCCTAAGGGTAGCAGC-3′); *PAT* (forward primer 5′-TGAGGGTGTTGTGGCTGGTA-3′; reverse primer 5′-TGTCCAATCGTAAGCGTTCCT-3′; probe (5′ FAM, 3′ NFQ MGB probe) 5′-TGCTTACGCTGGGCCCTGGAAG-3′). qRT–PCR was performed as follows: cDNA was synthesized from 2.5 μg of isolated RNA using the SuperScript III system (Invitrogen), diluted 1:5 with water and subjected to amplification reactions in a 20 μl volume containing 750 nM of each primer, 250 nM TaqMan probe, 25 ng of input total RNA in 2–5 μl volume and 1× TaqMan Gene Expression Master Mix (Applied Biosystems). Amplifications were performed on an iCycler (BIO-RAD) in a 96-well reaction plate as per the manufacturer's recommendations. Relative expressions of *IPK1* and *PAT* were calculated from ΔC_t values of *IPK1* or *PAT* and *EF1α* using the $2^{-\Delta\Delta C_t}$ method[40].

Plant cell culture. Embryogenic cell cultures of maize variety HiII[41] were generated, maintained and subjected to plasmid DNA delivery, selection of callus tissue and subsequent regeneration as described[42]. After fertilization, harvested seeds were dried and germinated directly in soil. Herbicide-tolerance selection was carried out by spraying 2-week-old seedlings with bialaphos (Liberty) at a dose equivalent of 280 g ha^{-1} or quizalofop (AssureII) at a dose equivalent of 35 g ha^{-1}.

Inorganic phosphate and phytate analysis. Individual seeds from selected plants were mechanically crushed and extracted with 2 N HCl using previously established methods with minor modifications[43,44]. Extracts were analysed by electrospray mass spectrometry using a minor modification of previous methods[45,46]. Peak areas were integrated for the parent ion of phytate (InsP6), inorganic phosphate (P_i) and intermediate inositol phosphate species. Results are reported by expressing the InsP6 and P_i as a fraction of each component normalized to the sum of the measured peak areas. Individual seed extracts were analysed in triplicate.

31. Isalan, M., Klug, A. & Choo, Y. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotechnol.* **19,** 656–660 (2001).
32. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature Biotechnol.* **26,** 702–708 (2008).
33. Maddaloni, M. *et al.* The sequence of the zein regulatory gene opaque-2 (O2) of *Zea mays*. *Nucleic Acids Res.* **17,** 7532 (1989).
34. Fang, J. *et al.* Stable antibody expression at therapeutic levels using the 2A peptide. *Nature Biotechnol.* **23,** 584–590 (2005).
35. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature Biotechnol.* **25,** 778–785 (2007).
36. Christensen, A. H., Sharrock, R. A. & Quail, P. H. Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. *Plant Mol. Biol.* **18,** 675–689 (1992).
37. McElroy, D., Zhang, W., Cao, J. & Wu, R. Isolation of an efficient actin promoter for use in rice transformation. *Plant Cell* **2,** 163–171 (1990).
38. Wohlleben, W. *et al.* Nucleotide sequence of the phosphinothricin N-acetyltransferase gene from *Streptomyces viridochromogenes* Tü494 and its expression in *Nicotiana tabacum*. *Gene* **70,** 25–37 (1988).
39. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: a Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, 1989).
40. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_t}$ method. *Methods* **25,** 402–408 (2001).
41. Armstrong, C., Green, C. & Phillips, R. Development and availability of germplasm with high type II culture formation response. *Maize Genet. Coop. News Lett.* **65,** 92–93 (1991).
42. Petolino, J. F., Hopkins, N. L., Kosegi, B. D. & Skokut, M. Whisker-mediated transformation of embryogenic callus of maize. *Plant Cell Rep.* **19,** 781–786 (2000).

43. Skoglund, E., Carlsson, N.-G. & Sandberg, A.-S. Determination of isomers of inositol mono- to hexaphosphates in selected foods and intestinal contents using high-performance ion chromatography. *J. Agric. Food Chem.* **45,** 431–436 (1997).
44. Stevenson-Paulik, J. *et al.* Generation of phytate-free seeds in *Arabidopsis* through disruption of inositol polyphosphate kinases. *Proc. Natl Acad. Sci. USA* **102,** 12612–12617 (2005).
45. Buscher, B. A. P., van der Hoeven, R. A. M., Tjaden, U. R., Andersson, E. & Van der Greef, J. Analysis of inositol phosphates and derivatives using capillary zone electrophoresis-mass spectrometry. *J. Chromatogr. A* **712,** 235–243 (1995).
46. Hsu, F.-F., Turk, J. & Gross, M. L. Structural distinction among inositol phosphate isomers using high-energy and low-energy collisional-activated dissociation tandem mass spectrometry with electrospray ionization. *J. Mass Spectrom.* **38,** 447–457 (2003).

Vol 459|21 May 2009|doi:10.1038/nature07845

LETTERS

High-frequency modification of plant genes using engineered zinc-finger nucleases

Jeffrey A. Townsend[1]*, David A. Wright[1]*, Ronnie J. Winfrey[1], Fengli Fu[1], Morgan L. Maeder[2,3], J. Keith Joung[2,3,4] & Daniel F. Voytas[5,6]

An efficient method for making directed DNA sequence modifications to plant genes (gene targeting) is at present lacking, thereby frustrating efforts to dissect plant gene function and engineer crop plants that better meet the world's burgeoning need for food, fibre and fuel. Zinc-finger nucleases (ZFNs)—enzymes engineered to create DNA double-strand breaks at specific loci—are potent stimulators of gene targeting[1,2]; for example, they can be used to precisely modify engineered reporter genes in plants[3,4]. Here we demonstrate high-frequency ZFN-stimulated gene targeting at endogenous plant genes, namely the tobacco acetolactate synthase genes (*ALS SuRA* and *SuRB*), for which specific mutations are known to confer resistance to imidazolinone and sulphonylurea herbicides[5]. Herbicide-resistance mutations were introduced into *SuR* loci by ZFN-mediated gene targeting at frequencies exceeding 2% of transformed cells for mutations as far as 1.3 kilobases from the ZFN cleavage site. More than 40% of recombinant plants had modifications in multiple *SuR* alleles. The observed high frequency of gene targeting indicates that it is now possible to efficiently make targeted sequence changes in endogenous plant genes.

ZFNs were engineered that recognize *SuR* loci, using publicly available resources provided by the Zinc Finger Consortium[6,7]. The Consortium-sponsored ZFN architecture uses two zinc-finger arrays (ZFAs), each with three zinc-fingers that collectively recognize a 9 base pair (bp) target site (Supplementary Fig. 1a). The ZFAs are fused to a FokI nuclease domain, and a 5–7 bp spacer separates the target sites for the two arrays, allowing the nuclease to dimerize and cleave within the spacer. ZFA engineering is most robust for G-rich sequences[8], and four such target sites were selected in *SuRB* for constructing ZFAs by modular assembly, namely the joining together of individual zinc-fingers with predetermined specificities (sites 815, 1071, 1853 and 1947) (Fig. 1a, Supplementary Table 1)[9]. Thirty-two ZFAs were constructed, and electrophoretic mobility shift assays identified three arrays with DNA binding activity, two of which bind half sites for the 815 target. This low success rate is consistent with previous findings that ZFAs constructed by modular assembly are often non-functional[8].

Oligomerized pool engineering (OPEN)—a method developed by the Consortium—uses genetic selections in bacteria to identify ZFA variants that recognize specific target sequences[6] (Supplementary Fig. 1b). ZFAs made by OPEN typically show higher activity than those made by modular assembly, probably because the process of selection accommodates context-dependent interactions among neighbouring zinc-fingers in the array[6,10,11]. OPEN was used to generate ZFNs for four sites (sites 865, 1853, 1947, 2163), including two that had been targeted by modular assembly (Supplementary Table 2). Functional left and right ZFAs were obtained for the 1853 target, for which modular assembly had failed, as well as for target 2163. The OPEN-derived ZFAs showed activity in bacterial two-hybrid assays[7], in which binding of ZFAs upstream of a *lacZ* reporter gene activates expression (Supplementary Table 2).

To test whether the ZFAs function as ZFNs, an assay was developed that measures ZFN activity in yeast (Supplementary Fig. 2a). This assay uses a *lacZ* reporter gene with a 125 bp internal DNA sequence duplication. The ZFN target site is cloned between the duplicated sequences, and cleavage of the target site creates a functional *lacZ* gene through repair of the break by single strand annealing. ZFN activity is assessed by quantitative measurements of β-galactosidase activity. The six ZFAs for the 815, 1853 and 2163 target sites functioned effectively as ZFNs (Fig. 2a). The 815 left and 1853 right arrays showed the most activity, comparable to activity observed with a ZFN designed from the well-characterized Zif268 ZFA[2].

ZFNs were tested against their endogenous targets in tobacco by measuring whether they create mutations by non-homologous end-joining (NHEJ) (Supplementary Fig. 2b). ZFN-encoding constructs were electroporated into tobacco protoplasts, and the relevant target sites in *SuRA* and *SuRB* were amplified by PCR and subjected to high-throughput pyrosequencing[12]. The fraction of unique sequence reads showing size-polymorphisms (consistent with imprecise repair by NHEJ) was normalized to controls (Fig. 2b, Supplementary Table 3). Mutation frequencies were significantly higher for ZFN 815 at both *SuR* loci. Interestingly, the highest mutation frequencies were not at the intended target in *SuRB*, but rather at the corresponding sequence in *SuRA*, which differs by two nucleotides. Not only does the ZFN 815 (which was created by modular assembly) lack specificity, but the higher level of mutagenesis at *SuRA* relative to *SurB* suggests that other factors, such as chromatin or DNA methylation, influence access of this enzyme to target sites. In contrast to ZFN 815, the OPEN-designed ZFN 1853 only showed enhanced mutagenesis at its intended *SuRB* target, which differs in sequence from the *SuRA* site by a single nucleotide. This suggests that the genetic selections used in OPEN yield ZFNs with high specificity. No enhancement of mutagenesis was observed with ZFN 2163.

To measure whether the engineered ZFNs could stimulate incorporation of specific DNA sequence changes at *SuR* loci by homologous recombination (HR) (Supplementary Fig. 2b), three donor templates were constructed, each with a missense mutation that confers resistance to one or more herbicides (P191A, chlorsulphuron; S647T, imazaquin; W568L, chlorsulphuron and imazaquin)[5,13] (Fig. 1a). Silent nucleotide changes were introduced into codons adjacent to each mutation to distinguish the donor template from the native locus and spontaneous

[1]Department of Genetics, Development & Cell Biology, Iowa State University, Ames, Iowa 50011, USA. [2]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [3]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [4]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Department of Genetics, Cell Biology & Development, [6]Center for Genome Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA.
*These authors contributed equally to this work.

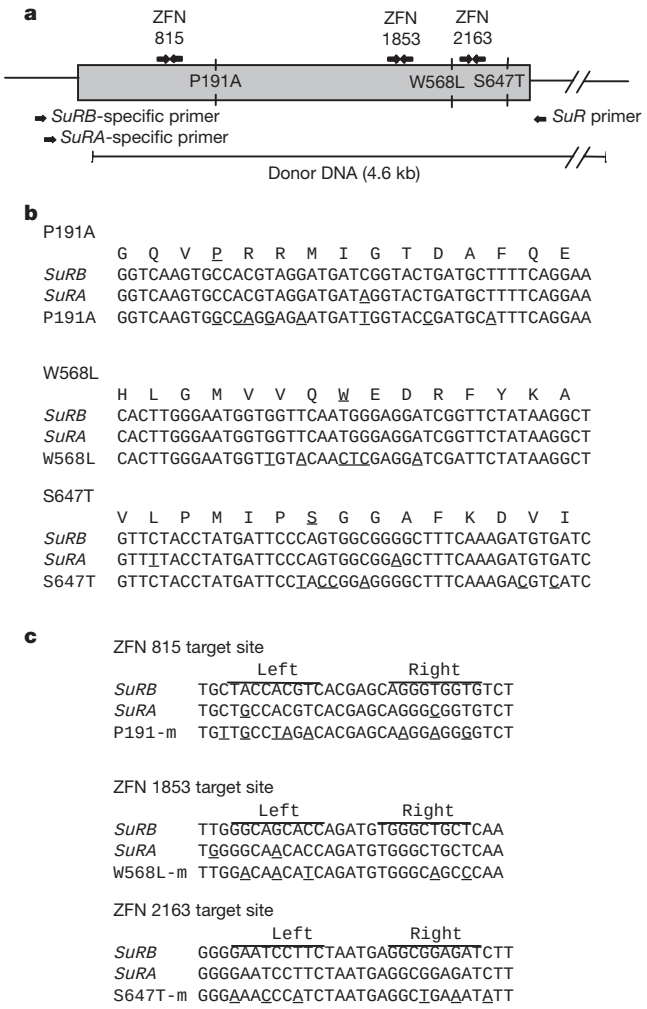


Figure 1 | The tobacco *SuRB* locus. a, This diagram is drawn to scale and annotated with ZFN sites, amino substitutions that confer herbicide resistance, PCR primers used to characterize recombinants, and the region used as a donor template. **b**, Sequences at the sites of introduced mutations. The targeted amino acid is underlined, as are sequences in *SuRA* that differ from *SuRB* and silent nucleotide changes in the donor template that distinguish recombinants from spontaneous mutants. **c**, ZFN target sites. 'Left' and 'right' denote bases recognized by each ZFA. Underlined bases are either *SuRA* sequences that differ from *SuRB* or mutated bases in the donor (bottom row) that prevent cleavage by the ZFN.

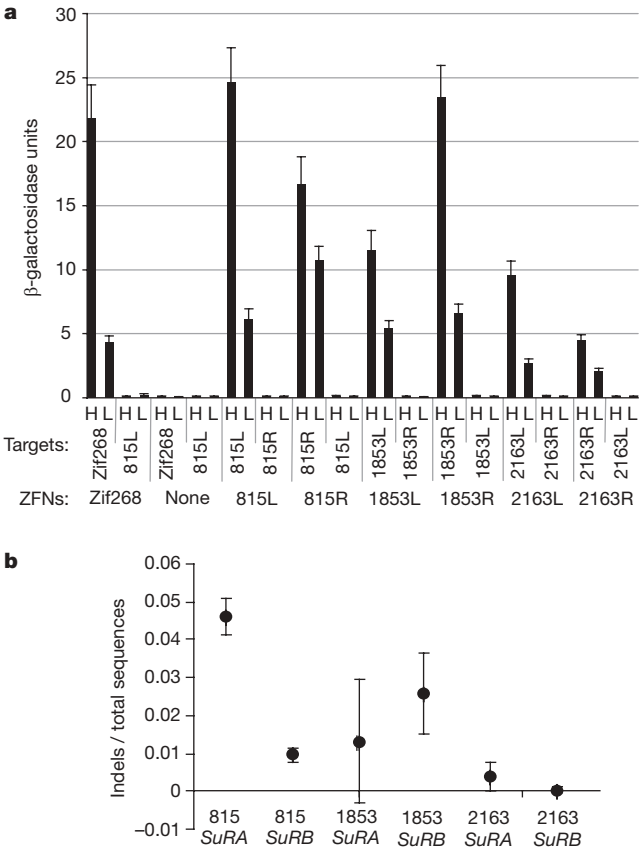


Figure 2 | Activity of engineered ZFAs and ZFNs. a, ZFAs as stimulators of recombination in yeast. Target sites for each ZFA are listed in vertical text below the chart; H, high-copy plasmid; L, low-copy plasmid. Error bars denote s.d.; $n = 3$. **b**, Engineered ZFNs as stimulators of mutagenesis by NHEJ in tobacco. ZFNs were expressed in protoplasts, and the *SuRA* and *SuRB* target sites were analysed by pyrosequencing. The number of sequences with insertions/deletions (indels) was divided by the total number of reads for a given target and normalized to a Zif268 control. Values above the *x* axis indicate a higher proportion of sequences with indels than the control. Error bars denote 95% confidence intervals; $n = 4$.

mutants from those generated by recombination (Fig. 1b). An additional set of donor templates was made in which the ZFN recognition sites were altered to prevent cleavage (Fig. 1c).

To test for gene targeting by HR, plasmids encoding the 815 ZFNs were electroporated into tobacco protoplasts with donor templates bearing the P191A, W568L or S647T mutations (Table 1, rows 1–3). The mean ZFN-induced herbicide resistance ranged from 5.3% for the P191A donor to 2.4% for the S647T donor. Both *SuRA* and *SuRB* were PCR-amplified from 12 randomly selected resistant calli derived from each treatment, using primers specific for the target locus (Fig. 1a). DNA sequence analysis revealed that in 9 of 12 lines generated with both the P191A and W568L donor template, resistance was due to HR (P191A: 5 in *SuRA* and 4 in *SuRB*; W568L: 4 at *SuRA*, 5 at *SuRB*) (Table 2, rows 1–3, Supplementary Table 4). With the S647T donor template, only 1 of the 12 herbicide resistant lines had evidence of HR (at *SuRA*) and nine were spontaneous *SuRB* mutants. For 8 of the 36

Table 1 | Gene targeting frequencies at *SuRA* and *SuRB*

Row	ZFN	Donor DNA*	Distance to mutation (bp)	Number herbicide resistant			Freq. ZFN-induced resistance‡ (%)	Freq. gene targeting§ (%)
				Neg. control	Pos. control	ZFN + donor†		
1	815	P191A	188	0	3,461	184	5.3	4.0
2	815	W568L	1,319	0	2,217	75	3.4	2.6
3	815	S647T	1,541	0	3,600	88	2.4	0.2
4	815	P191A-m	188	0	1,804	46	2.5	2.4
5	1853	W568L-m	281	0	1,577	10	0.6	0.5

* The letter m in column 3 denotes donors with mutated ZFN recognition sites.
† Values are the mean resistant calli obtained in three separate experiments.
‡ Values are the mean resistant calli divided by the number of transformed cells (based on data obtained from the positive control construct) and expressed as a percentage.
§ Values are the percentage of ZFN-induced herbicide resistance adjusted according to the frequency of HR as determined through molecular analyses of randomly sampled calli (Table 2).

Table 2 | Molecular basis for herbicide resistance in gene targeting experiments

Row	ZFN	Donor DNA	*SuRA*			*SuRB*		
			ZFN target	Mutation site		ZFN target	Mutation site	
			NHEJ indels/ alleles examined	HR events/ alleles examined	Spont. mutants/ alleles examined	NHEJ indels/ alleles examined	HR events/ alleles examined	Spont. mutants/ alleles examined
1	815	P191A	0/24	5/24	0/24	8/24	4/24	0/24
2	815	W568L	0/24	4/24	0/24	3/24	5/24	0/24
3	815	S647T	0/24	1/24	0/24	2/24	0/24	9/24
4	815	P191A-m	3/36	0/36	0/36	3/36	20/36	1/36
5	1853	W568L-m	0/22	0/22	1/22	2/22	9/22	1/22

NHEJ-induced mutations, HR events, and spontaneous (Spont.) mutations are expressed in terms of the number of alleles of *SuRA* or *SuRB* analysed (compiled from Supplementary Tables 4 and 5). Note that some plants sustained HR at more than one allele and that herbicide resistant somaclonal variants were recovered with no mutations in *SuRA* or *SuRB*. The letter m in column 3 denotes donors with mutated ZFN recognition sites.

resistant lines in the gene targeting experiments, no mutations were observed in *SuRA* or *SuRB*, and so the molecular basis for the resistance is unknown. This resistance could be due to genotypic and phenotypic variation (somoclonal variation) typically observed when plant cells are grown in culture[14]. Based on the number of recombinants recovered, the estimated gene targeting frequencies range from 4.0% for P191 to 0.2% for S647 (Table 1).

One surprising outcome of the above experiment was that gene targeting frequencies exceeding 2% were obtained at a distance more than 1.3 kilobases (kb) from the cleavage site. This suggests that plant genes can be modified even when DNA sequence composition precludes engineering ZFNs near the desired site of modification. The high frequencies of recombination observed at both *SuRA* and *SuRB* with ZFN 815 are consistent with the pyrosequencing data indicating that this enzyme cuts promiscuously at both targets (Fig. 2b). *SuRA* and *SuRB* differ at the nucleotide sequence level by 4% (ref. 5), and it is notable that high-efficiency gene targeting could be achieved at *SuRA* using the *SuRB*-derived donor template.

We next tested the ability of the 1853 and 2163 ZFNs to stimulate HR and incorporate amino acid sequence changes near their respective target sites. Donor templates were used with mutations in the ZFN target site that prevent cleavage. ZFN 815 was used as a control, and the mutated donor did not substantially alter the overall frequency of herbicide resistance or gene targeting (Table 1, compare rows 1 and 4). The mutated P191A donor template did, however, cause an increase in the proportion of gene targeting events at *SuRB* relative to *SuRA* (Table 2, compare rows 1 and 4). Why inability to cleave the donor template influences the outcome of recombination is unclear. For the 1853 ZFN, the mean number of herbicide resistant events at W568L (281 bp from the cut site) was 0.6% (Table 1, row 5), more than fivefold lower than gene targeting observed with ZFN 815 at much greater distances from the cut site. The 2163 ZFN yielded only three, non-targeted herbicide resistant calli in twelve separate experiments, two of which had mutations at sites in *SuRB* previously known to confer herbicide resistance[13] (Supplementary Table 5 and data not shown). The activity of all three ZFNs in HR parallels activities of these enzymes in the yeast and NHEJ assays (Fig. 2). Among the 47 herbicide resistant calli analysed in the various gene targeting experiments, 19 (40.4%) showed modifications at multiple *SuR* loci, including mutations introduced by NHEJ at the ZFN cleavage site (Supplementary Tables 4 and 5). This indicates that transformed cells that sustain a ZFN-induced modification often incur changes at multiple alleles. Ten plants were regenerated from herbicide resistant calli and shown to carry the *SuRA* and *SuRB* mutations (Supplementary Tables 4 and 5), indicating that ZFN-assisted gene targeting can be used to engineer genetically modified whole plants.

Based on the high frequency of gene targeting observed at *SuRA* and *SuRB*, we reasoned that populations of cells transformed with ZFNs might be screened directly for gene targeting events. To test this, protoplasts were transformed with plasmids encoding the 815 ZFN, the P191A donor template, as well as a neomycin phosphotransferase II (NPTII) gene that confers resistance to kanamycin. In this experiment, NPTII was used merely to identify cells that had been transformed. Approximately 1,000 transformed cells were transferred to media with herbicide, and two calli displayed herbicide resistance, both of which carried mutations introduced by the donor template (Supplementary Table 4). Although this experiment identified HR events at the *SuR* loci by examining herbicide resistance, its success suggests that screens, perhaps using high throughput DNA sequencing, could be used to identify recombinants among populations of transformants for any genetic modification introduced by recombination, regardless of whether a selection exists for its associated phenotype.

Current methods for modifying plant genomes are limited to decades-old methods of DNA transformation that lack precision and control over the outcome of the modified chromosome. Plant biologists have long sought a method to make directed mutations in plant genes with high efficiency, as evidenced here with the use of ZFNs. Gene targeting offers numerous opportunities for studying plant gene function, and it also enables biosynthetic pathways to be harnessed to better produce much-needed plant-derived products. The ability to engineer highly functional ZFNs using publicly available reagents and to recover HR-induced mutations even at considerable distances from the ZFN cleavage site demonstrates that targeted mutagenesis in plants is now practical.

METHODS SUMMARY

ZFA engineering and testing. Engineering of ZFAs by modular assembly or OPEN followed established protocols[6,7]. Assessments of ZFA function are described in greater detail in Supplementary Information and are summarized in Supplementary Table 6.

Plant cell transformation and culture. Tobacco protoplasts were prepared from aseptically grown plants and transformed by electroporation as previously described[4]. Before transformation, donor, ZFN-, and NPTII-encoding DNAs were linearized by digestion with BglI, AlwNI, or FspI, respectively. Electroporation experiments used 20 μg each of donor template and ZFN-encoding DNA. Chlorsulphuron (5 p.p.b.) and imazaquin (0.5 p.p.m.) were used for selection, the former for the P191A and W568L donor templates and the latter for the S647T donor. Transformation frequencies were assessed using *SuRB* genes with herbicide resistance mutations. Numbers of resistant calli were scored 30 days post treatment. All of the gene targeting experiments used obligate heterodimeric FokI domains fused to the left and right ZFAs[15,16]. Experiments with homodimeric (wild-type) FokI domains reduced plating efficiency relative to experiments with the heterodimeric domains, and no gene targeting events were recovered, consistent with high levels of ZFN-induced toxicity (Supplementary Table 7).

Characterization of recombinant plant material. DNA was isolated from calli with the PowerPlant DNA isolation kit (Mo Bio Laboratories). The entire coding regions of both *SuRA* and *SuRB* were PCR-amplified using primers specific to the 5′ end of each gene and a common 3′ primer (Fig. 1a). The Expand Long Template PCR System (Roche) was used to ensure fidelity of the PCR reactions and to minimize strand transfers during amplification. PCR products were sequenced in their entirety to identify mutations that confer herbicide resistance and modifications at the ZFN cut site.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 November 2008; accepted 30 January 2009.
Published online 29 April 2009.

1. Bibikova, M., Beumer, K., Trautman, J. K. & Carroll, D. Enhancing gene targeting with designed zinc finger nucleases. *Science* **300**, 764 (2003).

2. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300,** 763 (2003).
3. Lloyd, A., Plaisier, C. L., Carroll, D. & Drews, G. N. Targeted mutagenesis using zinc-finger nucleases in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **102,** 2232–2237 (2005).
4. Wright, D. A. *et al.* High-frequency homologous recombination in plants mediated by zinc-finger nucleases. *Plant J.* **44,** 693–705 (2005).
5. Lee, K. Y. *et al.* The molecular basis of sulfonylurea herbicide resistance in tobacco. *EMBO J.* **7,** 1241–1248 (1988).
6. Maeder, M. L. *et al.* Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol. Cell* **31,** 294–301 (2008).
7. Wright, D. A. *et al.* Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. *Nature Protocols* **1,** 1637–1652 (2006).
8. Ramirez, C. L. *et al.* Unexpected failure rates for modular assembly of engineered zinc fingers. *Nature Methods* **5,** 374–375 (2008).
9. Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F. III. Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA* **96,** 2758–2763 (1999).
10. Pruett-Miller, S. M., Connelly, J. P., Maeder, M. L., Joung, J. K. & Porteus, M. H. Comparison of zinc finger nucleases for use in gene targeting in mammalian cells. *Mol. Ther.* **16,** 707–717 (2008).
11. Cornu, T. I. *et al.* DNA-binding specificity is a major determinant of the activity and toxicity of zinc-finger nucleases. *Mol. Ther.* **16,** 352–358 (2008).
12. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281,** 363–365 (1998).
13. Tranel, P. & Wright, T. Resistance of weeds to ALS-inhibiting herbicides: What have we learned? *Weed Sci.* **50,** 700–712 (2002).
14. Kaeppler, S. M., Kaeppler, H. F. & Rhee, Y. Epigenetic aspects of somoclonal variation in plants. *Plant Mol. Biol.* **43,** 179–188 (2000).
15. Szczepek, M. *et al.* Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature Biotechnol.* **25,** 786–793 (2007).
16. Miller, J. C. *et al.* An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature Biotechnol.* **25,** 778–785 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Eichtinger for help in making ZFA reagents. This work was supported by grants to D.F.V. from the National Science Foundation and to J.K.J. from the National Institutes of Health and the Massachusetts General Hospital Department of Pathology.

Author Information DNA sequence of the *SuRB* locus has been deposited with GenBank under accession number FJ649655. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.F.V. (voytas@umn.edu).

METHODS

DNA donors. TAIL (thermal asymmetric interlaced)-PCR was used to amplify and clone a 7.8 kb genomic DNA fragment encompassing the *SuRB* locus. Oligonucleotide-directed mutagenesis was used to introduce mutations that confer herbicide resistance and silent nucleotide changes that distinguish HR events from spontaneous mutants (Fig. 1). Numbers specifying mutations and target sites are referenced with respect to GenBank accession X07645.

Three donor DNA pairs were constructed, each truncated 59 bp into the coding sequence. Truncation prevented herbicide resistance due to random integration of the donor DNA. pDW1963 and pDW1964 carry the P191A mutation; the latter has the 815 ZFN cut site mutated. pDW1927 and pDW1968 carry the W568L mutation; the latter has the 1853 ZFN cut site mutated. pDW1969 and pDW1972 have the S647T mutation; the latter has the 2163 ZFN site mutated. Donor DNAs are referred to by the mutation they carry (for example, W568L); 'm' indicates donors with mutated cut sites (for example, W568L-m).

Yeast assay. To test whether ZFAs function as ZFNs, a yeast-based recombination assay was developed similar to ones previously reported[17,18]. Our assay uses a derivative of the *Escherichia coli lacZ* gene with a 125 bp duplication of coding sequence. The sequence duplication flanks a *URA3* gene and a polylinker for cloning the target sequence being tested (Supplementary Fig. 2a). After cleavage of the target sequence by the ZFN, recombination between the *lacZ* sequence duplication results in loss of *URA3* (conferring 5-fluoroorotic acid resistance). Recombination also reconstitutes a functional *lacZ* gene, enabling quantitative measurements of enzyme activity that reflect frequencies of recombination.

A two-step process is used to construct yeast reporter plasmids. First, complementary oligonucleotides corresponding to the target site are cloned into the polylinker of plasmid pDW1666, which has the 3′ half of the *lacZ* gene. Second, a 2.5 kb NcoI to ApaI fragment from the pDW1666 derivative is inserted into the BspHI and ApaI sites of either plasmid pDW1714 or pDW1742. These latter plasmids contain the 5′ half of the *lacZ* gene, a *TRP1* marker, and either the 2μ origin of replication (pDW1714) or a yeast centromere (*CEN*) (pDW1742). Reporter plasmids are introduced into the yeast strain YPH500 (*MAT*α) and grown on synthetic complete medium lacking tryptophan and uracil (SC-W-U). The medium is adjusted to pH 7.0 and includes X-gal as previously described[17].

pDW1789 is used to express ZFNs in yeast from the *TEF1* promoter. ZFAs are cloned into a polylinker to generate FokI nuclease fusions. The expression plasmid has a *HIS3* gene and a *CEN*, and is introduced into YPH499 (*MAT*a). The transformed yeast strain is plated on SC medium lacking histidine (SC-H).

Quantitative β-galactosidase assays are performed by growing a white yeast colony carrying the target plasmid and a colony with the ZFN expression plasmid overnight at 30 °C in liquid SC-H or SC-W-U media, respectively. The cultures are adjusted to the same A_{600}, and 100 μl of each are added to 10 ml of yeast peptone dextrose (YPD) medium and cultured at 30 °C for 20–26 h. An aliquot of cells is harvested and quantitative β-galactosidase assays are performed as described[17]. Enzyme activity is normalized to cell number[7].

Plant ZFN expression vectors. ZFA::FokI fusions that recognize left and right half-sites of a given target are expressed in plant cells from a single plasmid. Expression plasmids are constructed in two steps: one ZFA is cloned as an XbaI/BamHI fragment into a polylinker in pDW1876; the second ZFA is cloned as an XbaI/BamHI fragment into a polylinker in pDW1895. The polylinkers in both pDW1876 and pDW1895 are flanked by CaMV 35S promoters and NOS transcriptional terminators. To make a dual ZFN expression plasmid, both plasmids are digested with ApaLI and AvrII. The fragment encoding the ZFN from the pDW1985 derivative is then cloned into the ApaLI to AvrII sites of the pDW1876 derivative. The above plasmids use the wild type FokI nuclease, and two variants (pDW2000 and pDW2001) express the obligate heterodimeric form of FokI to reduce cellular toxicity[18]. All the FokI nucleases were codon-optimized for expression in plants. pDW998 was used to identify transformed tobacco protoplasts. It carries a neomycin phosphotransferase gene (NPTII) expressed from a CaMV 35S promoter.

Tests of ZFN toxicity. To evaluate whether ZFN toxicity impacts the recovery of gene targeting events, aliquots of protoplasts were transformed with plasmids expressing the 815 ZFN with either the homodimeric or heterodimeric form of FokI nuclease (Methods Summary and Supplementary Table 7). The 815 ZFNs were separately electroporated with the three donor DNAs. Selection was imposed on the cells (Methods Summary) to identify possible recombinants. An aliquot of the culture was grown in the absence of selection, and plating efficiency was calculated by determining the number of calli that formed in the absence of selection relative to the total number of protoplasts treated.

Pyrosequencing to monitor NHEJ. Tobacco protoplasts were transformed with DNA encoding the 815, 1853 or 2163 ZFNs as described in the Methods Summary, except that 30 μg of ZFN-encoding DNA was used. For each of the target sites, four aliquots (experimental replicates) of protoplasts were transformed. An aliquot of protoplasts was also transformed with a ZFN derived from Zif268 to serve as a control. Forty-eight hours after transformation, DNA was prepared and the target sites were PCR-amplified with primers specific for given target sites and that amplify both *SuRA* and *SuRB*. The primers were bar-coded to distinguish the target sites, each of the experimental replicates, and the control. PCR products were gel purified and quantified using a Quant-iT PicoGreen kit (Invitrogen). Sequencing was carried out at the University of Iowa DNA Facility using the Roche 454 sequencer, yielding a total of 422,077 sequence reads. Sequences were parsed according to locus (*SuRA, SuRB*), target site (815, 1853, 2163), experimental replicate (Rep1 – Rep4), and control (Zif268). Analysis was limited to sequence reads showing size polymorphisms (indels) consistent with NHEJ-induced mutations. The frequency of indels was normalized to the Zif268 control as described in the text and the legend to Supplementary Table 3.

17. Epinat, J. C. *et al.* A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.* **31**, 2952–2962 (2003).
18. Doyon, Y. *et al.* Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. *Nature Biotechnol.* **26**, 702–708 (2008).

LETTERS

Crystal structure of the sodium–potassium pump at 2.4 Å resolution

Takehiro Shinoda[1], Haruo Ogawa[1], Flemming Cornelius[2] & Chikashi Toyoshima[1]

Sodium–potassium ATPase is an ATP-powered ion pump that establishes concentration gradients for Na^+ and K^+ ions across the plasma membrane in all animal cells by pumping Na^+ from the cytoplasm and K^+ from the extracellular medium[1,2]. Such gradients are used in many essential processes, notably for generating action potentials. Na^+,K^+-ATPase is a member of the P-type ATPases, which include sarcoplasmic reticulum Ca^{2+}-ATPase and gastric H^+,K^+-ATPase, among others, and is the target of cardiac glycosides. Here we describe a crystal structure of this important ion pump, from shark rectal glands, consisting of α- and β-subunits and a regulatory FXYD protein[3,4], all of which are highly homologous to human ones. The ATPase was fixed in a state analogous to $E2{\cdot}2K^+{\cdot}P_i$, in which the ATPase has a high affinity for K^+ and still binds P_i, as in the first crystal structure of pig kidney enzyme at 3.5 Å resolution[5]. Clearly visualized now at 2.4 Å resolution are coordination of K^+ and associated water molecules in the transmembrane binding sites and a phosphate analogue (MgF_4^{2-}) in the phosphorylation site. The crystal structure shows that the β-subunit has a critical role in K^+ binding (although its involvement has previously been suggested[6–8]) and explains, at least partially, why the homologous Ca^{2+}-ATPase counter-transports H^+ rather than K^+, despite the coordinating residues being almost identical.

The structure was determined using heavy-atom derivatives (Supplementary Table 1). Bound K^+ could be substituted by dialysis for Tl^+ and Rb^+, whose anomalous electron density maps showed three binding sites (two transmembrane (I and II) and one cytoplasmic (C); Figs 1b and 2d). The K^+ at site C is implicated in activation of dephosphorylation[9]. The atomic model (Fig. 1) was refined to an R_{free} value of 27.6% at 2.4 Å resolution (Supplementary Table 1) and for the α-subunit was, in general, very similar to that for pig kidney ATPase at 3.5 Å resolution[5]. The extracellular parts of the β-subunit and the FXYD protein have not been modelled previously.

Detailed comparison of the structure with the Ca^{2+}-ATPase in the corresponding state ($E2{\cdot}nH^+{\cdot}P_i$; Supplementary Fig. 1)[10] revealed important differences pertinent to those in function, although each of the three cytoplasmic domains[11] and transmembrane helices (M1–M10) are superimposable (Fig. 2). In Ca^{2+}-ATPase, the cytoplasmic headpiece has a compact configuration[10], stabilized, in part, by two hydrogen bonds between the actuator (A) and nucleotide (N) domains (Fig. 2d). In Na^+,K^+-ATPase, the A and N domains hardly interact, with only one salt bridge involving Glu 223 (A; corresponding to Val 185 in Ca^{2+}-ATPase) and Arg 551 (N; Arg 560) between them (Figs 1 and 2d). The N domain of Na^+,K^+-ATPase is ~22° further from the P domain than that of Ca^{2+}-ATPase (Fig. 2a), which is perhaps pertinent to the much faster ATPase turnover (~200 s^{-1} versus ~30 s^{-1}).

As the ATP-binding site is located just above this salt bridge, the ATP molecule in $E2{\cdot}ATP(TG)$ of Ca^{2+}-ATPase can be docked at the equivalent position (Fig. 2d). With ATP here, Arg 551 is likely to form a salt bridge to the β-phosphate. In fact, the Arg551→Gln mutation abolishes ATP binding[12]. Arg 560, the corresponding residue in Ca^{2+}-ATPase, interacts with the β-phosphate[13,14] and is a key residue in ATP binding[15]. As both ADP and ATP hinder formation of the complex with MgF_4^{2-} (ref. 16), it is likely that nucleotides disrupt the Glu 223–Arg 551 salt bridge, aiding dissociation of the A and N domains. If this salt bridge were maintained in $E2{\cdot}2K^+$, the well-known accelerating effect of ADP and ATP on K^+ release[2] would be readily explained. Such acceleration effects of ATP, though to a lesser degree, are also known for Ca^{2+}-ATPase (see, for example, ref. 17), and Arg 489 (N) appears to play a similar role together with Asp 203 (A)[18].

In the transmembrane region, the unwinding of M5 (at Asn 783) and M7 (at Gly 855) in Na^+,K^+-ATPase stands out (Fig. 1c). The segment of M7 near the cytoplasmic surface (M7′) has a distinct kink of ~18° (Fig. 2a), which appears to have central importance in K^+ binding. At 2.4 Å resolution, the coordination geometry of K^+ and associated water molecules is evident. Site I is made essentially of only five oxygen atoms, contributed by one main chain (Thr 779), three side chains (Ser 782, Asn 783 and Asp 811) and one water molecule (Fig. 3). The closest oxygen atom of Glu 786 (Glu 771 in Ca^{2+}-ATPase) is 3.9 Å away and does not contribute. Site I has a valence[19] of 1.06 (Supplementary Table 2), and is well qualified to be a high-affinity K^+ site (ideal valence, 1.0).

The site-II K^+ is shifted by 1.3 Å to the extracellular side in comparison with the site-I K^+ (Fig. 3b). It is coordinated by three main-chain carbonyls (Val 329, Ala 330 and Val 332), three or four side-chain oxygen atoms (Asn 783, Glu 786, Asp 811 and possibly Glu 334) and no water molecules (Fig. 3). Although the coordination number is high (six or seven), the sum of partial valences is lower than for site I, being 0.64 (Supplementary Table 2). Glu 334 on M4 (equivalent to Glu 309 in Ca^{2+}-ATPase) is structurally a very important residue, consistent with mutagenesis results[20–22], but marginally qualifies as a K^+-binding residue with the carboxyl 3.3 Å away (partial valence, 0.03). Thus, the two K^+ sites have different coordination characteristics but coordination geometry is rather distorted at both sites. The sum of partial valences[19] calculated for Na^+ is 0.62 at site I and 0.54 at site II (Supplementary Table 2), explaining the lower affinity for Na^+ in this state.

Unexpectedly, Asp 815, which is equivalent to Asp 800 in Ca^{2+}-ATPase and therefore expected to play a pivotal role in Na^+ binding[23], does not coordinate K^+, although it fixes a coordinating water at site I (Fig. 3). Gln 930 on M8 (equivalent to Glu 908 in Ca^{2+}-ATPase) is far from the bound K^+, but appears to stabilize the Asp 815 side chain, which is also stabilized by the Tyr 778 hydroxyl. Thus, they might contribute to adjust the affinity for K^+.

A possibly more unexpected observation is that the two bound K^+ ions (ionic radius, 1.35 Å) are juxtaposed only 4.1 Å apart, with no

[1]Institute of Molecular and Cellular Biosciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-0032, Japan. [2]Department of Physiology and Biophysics, University of Aarhus, DK-8000 Aarhus C, Denmark.

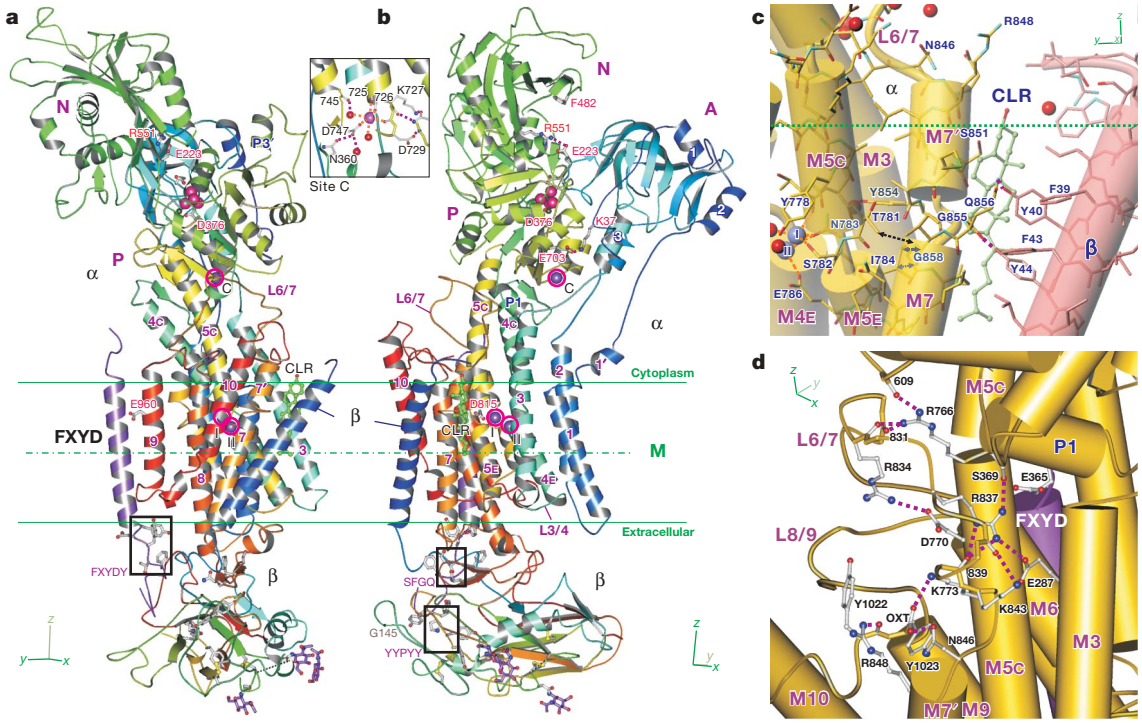


Figure 1 | Architecture of Na^+,K^+-ATPase with bound MgF_4^{2-} and K^+. **a**, **b**, Ribbon diagrams. **c**, **d**, Details around the unwound part of M7 (**c**) and the L6/7 loop (**d**). In **a** and **b**, the colour changes gradually between the amino-terminal (blue) and carboxy-terminal (red) ends for the α- and β-subunits. The FXYD protein is in purple. Circled purple spheres represent bound K^+ ions. A cholesterol molecule (CLR), sugars and several key residues are depicted in ball-and-stick style. The *x* and *y* axes respectively correspond to the *a* and *b* axes of the crystals. The membrane plane appears to lie at an inclination of ~4° from the *a*–*b* plane (Supplementary Fig. 2b). Inset, detail at the K^+-binding site in the phosphorylation (P) domain (site C); the coordination geometry is largely different from that in the pig kidney model[9]. The horizontal green lines show the approximate boundaries of the hydrophobic core of the lipid bilayer. Common motifs in the extracellular region are boxed. Hydrogen bonds (broken purple lines), K^+ coordination (broken orange lines) and van der Waals contacts (broken black arrows) are also shown. The dotted line in the β-subunit shows the part (nine residues) for which the atomic model was not built (**a**). The amino-terminal 31 residues of the α-subunit and 27 residues of the β-subunit and the carboxy-terminal 33 residues of FXYD10, all in the cytoplasm, could not be modelled. M, membrane; P1, P3′, α-helices in the P domain; OXT, terminal carboxyl oxygen.

intervening oxygen atom. Two side-chain oxygen atoms of Asp 811 and Asn 783 are shared by the two K^+ ions, but there is no carboxyl group bridging them. This is in marked contrast with the Ca^{2+}-binding sites in Ca^{2+}-ATPase[11], in which Asp 800 bridges the two Ca^{2+} ions, and no oxygen atom is shared by them. The short distance between the two K^+ ions and the orientation of Asp 811 are related to the binding cavity being highly confined.

Homology modelling based on Ca^{2+}-ATPase crystal structures[23] correctly predicted many features of the K^+-binding sites, indicating the similarity of these two pumps. For instance, according to the predictions a larger binding cavity for K^+ is created by bowing of M5 towards M1, which removes Glu 786 from site I and brings in the smaller Ser 782 instead, and rotation of M6 brings in Asp 811 and removes Thr 814 from site I. However, perhaps the most striking observation is that the positions of the coordinating residues, and even the conformations of the side chains, are virtually identical, except for that of Asn 783 (Fig. 3). Yet, whereas the K^+-binding sites in Na^+,K^+-ATPase accommodate other monovalent alkali cations including Na^+ and H^+ (see, for example, ref. 1), Ca^{2+}-ATPase binds only H^+ for counter-transport[24].

The crystal structures explain this. It is noteworthy that the K^+-binding sites are fairly confined and offset in comparison with the Ca^{2+} sites, very close to the M4 and M5 helices (Fig. 3a), despite K^+ being substantially larger than Ca^{2+} (1.35 Å versus 0.99 Å). This is a natural consequence of employing a main-chain carbonyl for coordinating K^+ at site I, rather than, for instance, the Asp 815 carboxyl as predicted[23]. In fact, usage of a main-chain carbonyl appears to be the key in defining ion selectivity.

A marked difference from Ca^{2+}-ATPase is the presence of Pro 785 on the M5 helix of Na^+,K^+-ATPase (instead of Gly 770 in Ca^{2+}-ATPase). Because of this Pro, M5 of Na^+,K^+-ATPase is unwound and kinked (~18°) such that the main-chain oxygen of Thr 779 becomes available for K^+ coordination (Fig. 3b). M5 of Ca^{2+}-ATPase is similarly kinked (Figs 2a and 3b), but is a continuous helix, creating little extra space. Reflecting this, the main-chain positions of Asn 783 and the corresponding Asn 768 are critically different. The amide of Asn 768 in Ca^{2+}-ATPase comes too close to the site-I K^+ (3.0 Å versus 3.8 Å; Fig. 3b), and the side chain too close to the site-II K^+, to allow K^+ binding. In Na^+,K^+-ATPase, the side chain of Asn 783, the corresponding residue, has a distinct conformation fixed by a hydrogen bond with Tyr 854, made possible by the kink of M7′ (Figs 2c and 3a). The cytoplasmic end of M7′ is positioned by a rigid L6/7 loop, critically linked to M3 and M5 with hydrogen bonds involving Arg 766 and Arg 837 (Fig. 1d), known as loci of familial hemiplegic migraine[25], and Glu287, a locus of rapid-onset dystonia-parkinsonism[26].

In Ca^{2+}-ATPase, the M5 helix straightens[18] during the $E2P \rightarrow E2 \rightarrow E1{\cdot}2Ca^{2+}$ transitions. If a similar movement takes place in Na^+,K^+-ATPase, straightening of M5 will push M7′ towards M10 (Fig. 2e), which in turn will cause large-scale structural changes involving P1, M3 and the carboxy-terminal segment of the α-subunit as well as the β-subunit, owing to the extensive hydrogen-bonding network (Fig. 1d). Thus, helices M7–M10, which looked like a mere membrane anchor in Ca^{2+}-ATPase, appear to have a dynamic role in Na^+,K^+-ATPase and might be used to change the position of the cytoplasmic domain of the β-subunit. We expect that the Asn 783 side chain will change its position to interact with the smaller Na^+ with the help of Pro 785, as the interactions between M5 and M7′, including hydrogen bonding between Asn 783 and Tyr 854, will break.

Thus, unwinding of M7 appears to be of central importance in K^+ binding. This unwinding is stabilized by hydrogen bonding of Tyr 44

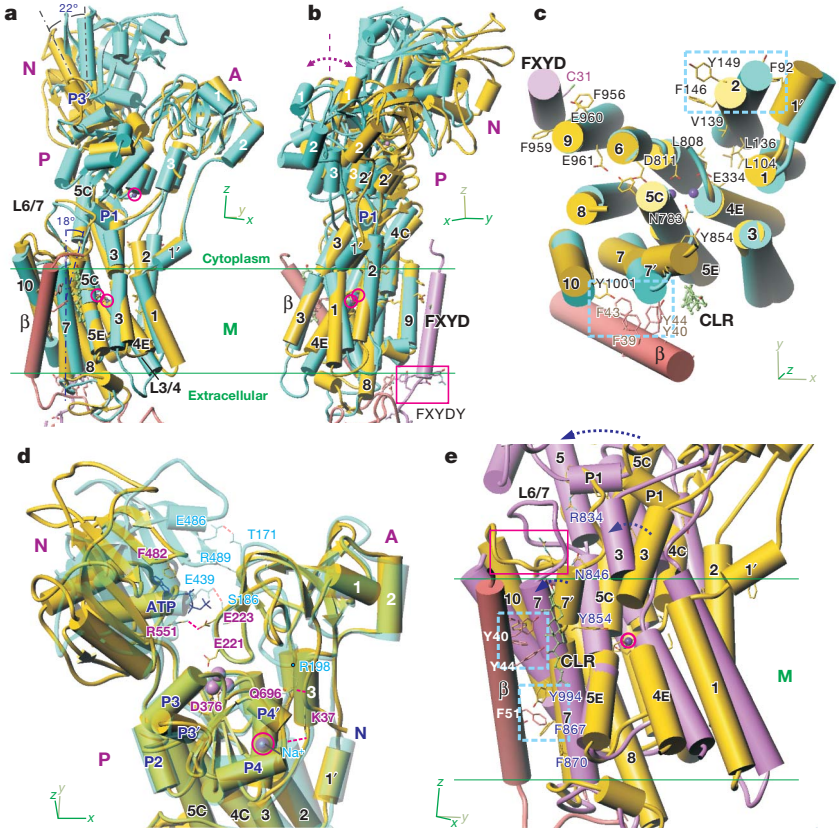


Figure 2 | Superimposition of the crystal structure of Na^+,K^+-ATPase on that of Ca^{2+}-ATPase. The structures are aligned with the six metal-coordinating residues[11] (**a**–**c**, **e**) or the P domain (**d**). In **a**–**d**, the E2·MgF_4^{2-}(TG) form of Ca^{2+}-ATPase (Protein Data Bank accession number, 1WPG) is superimposed; in **e**, the E1·$2Ca^{2+}$ form. The α-subunit of Na^+,K^+-ATPase is always in yellow, and the β-subunit in brown; the E2·MgF_4^{2-}(TG) form of Ca^{2+}-ATPase is in lime, the E1·$2Ca^{2+}$ form in violet. Although M3 is a continuous helix in both ATPases, it is represented by two cylinders. M5 and M7 are continuous helices in Ca^{2+}-ATPase but contain an unwound part in Na^+,K^+-ATPase; they are represented by three and two cylinders, respectively. In **e**, the transmembrane part of M3 is removed. Bound K^+ ions are represented by purple spheres in **a**–**c** and **e** (and marked with red circles). In **d**, MgF_4^{2-} and K^+ at site C (together with the corresponding Na^+ (cyan) in Ca^{2+}-ATPase) are shown in space-fill style, and cholesterol (CLR) in ball-and-stick style; an ATP taken from the E2·ATP(TG) form of Ca^{2+}-ATPase (Protein Data Bank accession number, 2DQS) is depicted using black sticks. The arrows in **e** indicate the expected movements in transition to the E1·$3Na^+$ state. The FXYDY motif (red box in **b**), the carboxy-terminal 10 residues (red box in **e**), and clusters of aromatic residues (cyan boxes in **c** and **e**) are also shown. TG, thapsigargin.

of the β-subunit to the Gly 855 carbonyl (Fig. 1c), which otherwise would be exposed to the hydrophobic core of the bilayer. Therefore, Ca^{2+}-ATPase cannot bind K^+ partly because it does not have a β-subunit. We note that there is a cholesterol molecule here (Fig. 1c and Supplementary Fig. 3), which appears to shield the unwound part of M7 from the bulk lipid. Its presence here is potentially related to a strong dependence of the Na^+,K^+-ATPase activity on cholesterol[27,28]. This cholesterol molecule, carried through from the native tissue, occupies the position in which a phospholipid head group was previously located[5].

The β-subunit is a single spanning membrane protein important in targeting and stabilization, but also affects the ion-binding and transport properties[8]. Its extracellular domain is large and rich in aromatic residues (Supplementary Fig. 4) and is glycosylated. We identified sugar residues at two of four potential sites (Fig. 1). The transmembrane helix of the β-subunit runs rather detached from those of the α-subunit and is inclined by ~32° from the membrane normal, nearly parallel to αM7 (Fig. 2). However, it forms four hydrogen bonds and numerous contacts with two transmembrane helices of the α-subunit, primarily using two clusters of aromatic residues. The first cluster, at about the level of the cholesterol ring, is formed from four residues (Phe 39–Tyr 44) of the β-subunit and one (Tyr 1001) from M10 of the α-subunit (Fig. 2c). Both Tyr 40 and

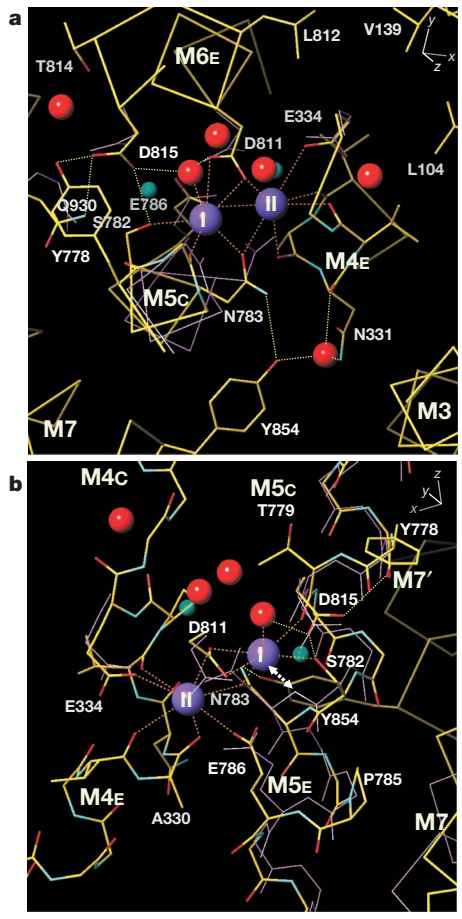


Figure 3 | Transmembrane K^+-binding sites. **a**, View from the cytoplasmic side; **b**, view approximately parallel to the membrane. Main chains of the M5 and M7 helices and side chains of the Ca^{2+}-coordinating residues of Ca^{2+}-ATPase (violet lines) are superimposed. The two structures are aligned with the metal-coordinating residues. Purple spheres represent bound K^+ ions, and red spheres water molecules; small cyan spheres (transparent) show the positions of bound Ca^{2+} ions in the E1·$2Ca^{2+}$ state of Ca^{2+}-ATPase. K^+ coordination (broken orange lines), hydrogen bonds (broken green lines) and a steric clash expected between K^+ at site I and the Asn 768 amide in Ca^{2+}-ATPase (corresponding to Asn 783 in Na^+,K^+-ATPase; white arrows in **b**) are also depicted. A stereo version of this figure is presented as Supplementary Fig. 6.

Tyr 44 interact with αM7 and cholesterol (Fig. 1c). They are highly conserved and the mutagenesis changes the apparent K^+ affinity[7]. The other cluster is in the extracellular leaflet of the bilayer, to which seven aromatic residues on four helices participate (Fig. 2e).

On the extracellular side, the carboxy-terminal part of the L7/8 loop (Asp 892–Gln 910) is the primary interaction site, in which a consensus sequence ^{901}SFGQ, proposed as a key interaction site[29], is located (Figs 1 and 4). The well-conserved Tyr 247 in the ^{244}YYPYY motif of the β-subunit fixes βArg 183 to mediate a salt bridge with αGlu 899, which in turn forms a salt bridge with βLys 250 (Fig. 4). Such complex interactions between the β- and α-subunits should explain, at least partly, previous results that implicate the β-subunit in modulation of cation transport[6,8].

Na^+, K^+-ATPase from shark rectal glands contains an accessory regulatory protein, FXYD10[3]. The transmembrane part runs approximately perpendicular to the membrane (Fig. 1a), interacting almost exclusively with the outside of αM9. Two conserved Gly residues, in particular Gly 34, are evidently important for this (Supplementary Fig. 5). Likely hydrogen bonds are found only between Cys 31 and αGlu 960. As the αGlu 960 carboxyl will be exposed to the hydrophobic part of the membrane if FXYD is absent (Fig. 2c), it will have to find a hydrogen-bond partner inside the α-subunit. Hence, it is conceivable that FXYD10 plays an important regulatory role through interactions with αGlu 960.

The role of the signature motif FXYD appears clear (Fig. 4). The first residue, Phe 12, anchors the segment to the β-subunit and the last residue, Asp 15, caps the helix. The third residue, Tyr 14, and the subsequent Tyr 16 form a cluster of aromatic residues with βTyr 69 and αTrp 987 to sandwich the β-subunit, together with the α-subunit. Glu 10/Arg 11 and Tyr 16 mediate a complex hydrogen-bonding network involving both subunits. Thus, functionally the motif should be considered to be FXYDY/W. The last Tyr is substituted for Trp in FXYD3 and FXYD4, but its indole ring will function similarly as the hydrogen donor to βAsp 71.

Previous studies located FXYD in the groove surrounded by M2, M6 and M9, following the model proposed for phospholamban[30], a regulatory protein for Ca^{2+}-ATPase that alters Ca^{2+} affinity presumably by interfering with the movement of M2. Three residues on M2 and another on M1 form an aromatic cluster there (Fig. 2c), apparently preventing access of a transmembrane helix.

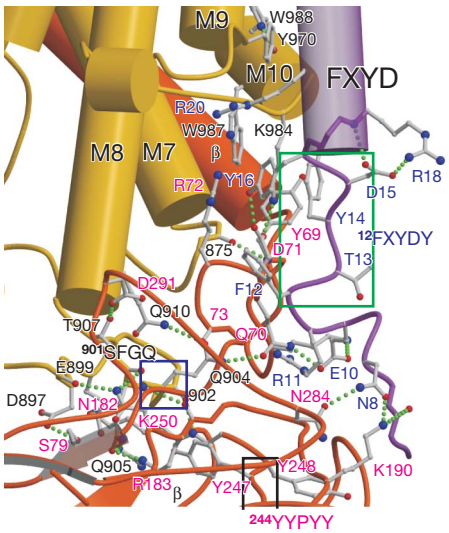


Figure 4 | Interactions among the α- and β- subunits and the FXYD protein at the extracellular surface of the membrane. View approximately parallel to the membrane. Conserved motifs are identified with boxes. Dotted lines represent hydrogen bonds. Four regions of the β-subunit contribute to interactions with the α-subunit: Tyr 69–Tyr 83 (also interacting with FXYD), Asn 182–Ile 185, Tyr 247–Lys 250 and Lys 290–Asp 291. Key residues appear to be Glu 899, Gln 904 and Gln 905 on the α-subunit and Asn 182 and Arg 183 on the β-subunit. Gln 905 is a locus of familial hemiplegic migraine[25]. A stereo version of this figure is presented as Supplementary Fig. 7.

METHODS SUMMARY

Na^+, K^+-ATPase was isolated from shark rectal glands and purified by mild deoxycholate extraction followed by differential centrifugation. This preparation contained the α- and β-subunits and an equimolar amount of FXYD10 (ref. 3). Crystals were grown by the dialysis method, supplemented with phosphatidylcholine, in the presence of MgF_4^{2-} and K^+ or its congener. Molecular replacement starting from the published model[5] was successful for resolving the α-subunit and FXYD. Multiple isomorphous replacement was performed to model the β-subunit.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 July 2008; accepted 26 February 2009.

1. Albers, R. W. Biochemical aspects of active transport. *Annu. Rev. Biochem.* **36,** 727–756 (1967).
2. Post, R. L., Hegyvary, C. & Kume, S. Activation by adenosine triphosphate in the phosphorylation kinetics of sodium and potassium ion transport adenosine triphosphatase. *J. Biol. Chem.* **247,** 6530–6540 (1972).
3. Mahmmoud, Y. A., Vorum, H. & Cornelius, F. Identification of a phospholemman-like protein from shark rectal glands. Evidence for indirect regulation of Na,K-ATPase by protein kinase c via a novel member of the FXYDY family. *J. Biol. Chem.* **275,** 35969–35977 (2000).
4. Garty, H. & Karlish, S. J. Role of FXYD proteins in ion transport. *Annu. Rev. Physiol.* **68,** 431–459 (2006).
5. Morth, J. P. *et al.* Crystal structure of the sodium–potassium pump. *Nature* **450,** 1043–1049 (2007).
6. Lutsenko, S. & Kaplan, J. H. An essential role for the extracellular domain of the Na,K-ATPase β-subunit in cation occlusion. *Biochemistry* **32,** 6737–6743 (1993).
7. Hasler, U., Crambert, G., Horisberger, J. D. & Geering, K. Structural and functional features of the transmembrane domain of the Na,K-ATPase β subunit revealed by tryptophan scanning. *J. Biol. Chem.* **276,** 16356–16364 (2001).
8. Geering, K. The functional role of β subunits in oligomeric P-type ATPases. *J. Bioenerg. Biomembr.* **33,** 425–438 (2001).
9. Schack, V. R. *et al.* Identification and function of a cytoplasmic K^+ site of the Na^+, K^+ -ATPase. *J. Biol. Chem.* **283,** 27982–27990 (2008).
10. Toyoshima, C., Nomura, H. & Tsuda, T. Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature* **432,** 361–368 (2004).
11. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405,** 647–655 (2000).
12. Jacobsen, M. D., Pedersen, P. A. & Jorgensen, P. L. Importance of Na,K-ATPase residue alpha 1-Arg544 in the segment Arg544-Asp567 for high-affinity binding of ATP, ADP, or MgATP. *Biochemistry* **41,** 1451–1456 (2002).
13. Toyoshima, C. & Mizutani, T. Crystal structure of the calcium pump with a bound ATP analogue. *Nature* **430,** 529–535 (2004).
14. Sørensen, T. L., Møller, J. V. & Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science* **304,** 1672–1675 (2004).
15. Clausen, J. D., McIntosh, D. B., Vilsen, B., Woolley, D. G. & Andersen, J. P. Importance of conserved N-domain residues Thr441, Glu442, Lys515, Arg560, and Leu562 of sarcoplasmic reticulum Ca^{2+}-ATPase for MgATP binding and subsequent catalytic steps. Plasticity of the nucleotide-binding site. *J. Biol. Chem.* **278,** 20245–20258 (2003).
16. Murphy, A. J. & Hoover, J. C. Inhibition of the Na,K-ATPase by fluoride. Parallels with its inhibition of the sarcoplasmic reticulum CaATPase. *J. Biol. Chem.* **267,** 16995–17000 (1992).
17. Jensen, A. M., Sørensen, T. L., Olesen, C., Møller, J. V. & Nissen, P. Modulatory and catalytic modes of ATP binding by the calcium pump. *EMBO J.* **25,** 2305–2314 (2006).
18. Toyoshima, C. & Nomura, H. Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature* **418,** 605–611 (2002).
19. Brown, I. D. & Wu, K. K. Empirical parameters for calculating cation-oxygen bond valences. *Acta Crystallogr. B* **32,** 1957–1959 (1976).
20. Jewell-Motz, E. A. & Lingrel, J. B. Site-directed mutagenesis of the Na,K-ATPase: consequences of substitutions of negatively-charged amino acids localized in the transmembrane domains. *Biochemistry* **32,** 13523–13530 (1993).
21. Nielsen, J. M., Pedersen, P. A., Karlish, S. J. & Jorgensen, P. L. Importance of intramembrane carboxylic acids for occlusion of K^+ ions at equilibrium in renal Na,K-ATPase. *Biochemistry* **37,** 1961–1968 (1998).
22. Vilsen, B. & Andersen, J. P. Mutation to the glutamate in the fourth membrane segment of Na^+,K^+-ATPase and Ca^{2+}-ATPase affects cation binding from both sides of the membrane and destabilizes the occluded enzyme forms. *Biochemistry* **37,** 10961–10971 (1998).

23. Ogawa, H. & Toyoshima, C. Homology modeling of the cation binding sites of Na^+K^+-ATPase. *Proc. Natl Acad. Sci. USA* **99,** 15977–15982 (2002).
24. Ueno, T. & Sekine, T. Study on calcium transport by sarcoplasmic reticulum vesicles using fluorescence probes. *J. Biochem.* **84,** 787–794 (1978).
25. Pietrobon, D. Familial hemiplegic migraine. *Neurotherapeutics* **4,** 274–284 (2007).
26. de Carvalho Aguiar, P. *et al.* Mutations in the Na^+/K^+ -ATPase α_3 gene ATP1A3 are associated with rapid-onset dystonia parkinsonism. *Neuron* **43,** 169–175 (2004).
27. Cornelius, F., Turner, N. & Christensen, H. R. Modulation of Na,K-ATPase by phospholipids and cholesterol. II. Steady-state and presteady-state kinetics. *Biochemistry* **42,** 8541–8549 (2003).
28. Sotomayor, C. P., Aguilar, L. F., Cuevas, F. J., Helms, M. K. & Jameson, D. M. Modulation of pig kidney Na^+/K^+-ATPase activity by cholesterol: role of hydration. *Biochemistry* **39,** 10928–10935 (2000).
29. Colonna, T. E., Huynh, L. & Fambrough, D. M. Subunit interactions in the Na,K-ATPase explored with the yeast two-hybrid system. *J. Biol. Chem.* **272,** 12366–12372 (1997).
30. Toyoshima, C. *et al.* Modeling of the inhibitory interaction of phospholamban with the Ca^{2+} ATPase. *Proc. Natl Acad. Sci. USA* **100,** 467–472 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Kawamoto and N. Shimizu for their help in data collection at BL41XU, SPring-8, and T. Tsuda for many aspects of this work. We are grateful to D. B. McIntosh for his help in improving the manuscript and H. R.Z. Christensen for technical assistance. Thanks are also due to G. Cramb for sharing sequencing results of the β-subunit with us before publication. This work was supported by a Specially Promoted Project Grant from the Ministry of Education, Culture, Sports, Science and Technology of Japan, to C.T., and grants from the Danish Medical Research Council, to F.C.

Author Information Atomic coordinates and structure factors for the structure reported in this work have been deposited in the Protein Data Bank under accession number 2ZXE. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.T. (ct@iam.u-tokyo.ac.jp).

METHODS

Preparation of shark Na^+, K^+-ATPase. Crude membrane fractions (microsomes) from the rectal gland of the shark *Squalus acanthias* were prepared by homogenization followed by washing and isolation by centrifugation in 30 mM histidine, 1 mM EDTA, 0.25 M sucrose, pH 6.8. The microsomal preparation was subsequently purified by sucrose flotation[31]. The microsomes were diluted to 40% sucrose and layered on top of 60% sucrose followed by layers of 35% sucrose and histidine/EDTA buffer without sucrose. After centrifugation at 96,000*g* for 2 h at 4 °C, the bands at the 0/35% and 35/40% interfaces were collected, washed and resuspended in the histidine/EDTA buffer with 0.25 M sucrose.

The purified microsomes were washed with ~0.15% deoxycholate to remove extrinsic proteins and to open sealed vesicles. Then a purified membrane preparation was obtained by differential centrifugation essentially as described previously[32]. The preparation was suspended in the histidine/EDTA buffer with 25% glycerol and kept at −20 or −80 °C until use. The preparation showed a turnover number of 200 s^{-1} at 37 °C.

Crystallization. Solubilised ATPase with octaethyleneglycol mono-*n*-dodecylether ($C_{12}E_8$) was mixed with a buffer supplemented with exogenous phosphatidylcholine, consisting of 100 mM KCl, 4 mM $MgCl_2$, 8 mM KF, 5 mM glutathione, 15 mg ml^{-1} $C_{12}E_8$, 20% (w/v) glycerol and 20 mM Tris buffer/HCl, pH 7.0, so that the final concentrations of ATPase and phosphatidylcholine were 2.5 mg ml^{-1} and 2.1 mg ml^{-1}, respectively. The solution was put in dialysis buttons and dialysed against a buffer consisting of 18% (w/v) polyethyleneglycol 3000, 25% glycerol (w/v), 5% (v/v) 2-methyl-2,4-pentanediol, 100 mM potassium acetate, 10 mM KCl, 4 mM $MgCl_2$, 4 mM KF, 0.1 mM EGTA, 10 mM glutathione, 2 μg ml^{-1} 2,6-di-*t*-butyl-*p*-cresol, 20 mM MES/Tris buffer, pH 7.0, at 25 °C for 1–2 months. Before flash freezing with cold nitrogen gas, the specimen was dialysed overnight against the same buffer but containing 40% (w/v) polyethyleneglycol 3000. This treatment shrank the *a*-axis dimension and improved the resolution.

Structure determination and analysis. All the diffraction data were collected at BL41XU of SPring-8, Japan, at 100 K using ADSC Q315 and Rayonix E255HE charge-coupled-device detectors. The wavelength used was 0.9 Å for the native data sets and those corresponding to respective absorption peaks for the derivatives. DENZO and SCALEPACK[33] were used to process diffraction data. The crystals belonged to the *C*2 space group with unit-cell parameters of $a = 223.8$ Å, $b = 50.9$ Å, $c = 163.8$ Å and $\beta = 105.1°$. The *a*-axis dimension varied between 219 and 226 Å. MLPHARE in the CCP4 package[34] was used for MIRAS (multiple isomorphous replacement with anomalous scattering) phasing. The mean figure of merit was 0.55 using MIRAS and 0.98 after density modification with CNS[35]. Molecular replacement and refinement were carried out with CNS, and finally with REFMAC[36]. Fractions of residues (1,134 altogether, excluding Gly and Pro) in the most favoured, additionally allowed, generously allowed and disallowed regions of the Ramachandran plot are 84.0%, 14.9%, 1.2% and 0.0%, respectively, according to PROCHECK[37]. Hydrogen bonds were identified with HBPLUS[38]. Structural figures were prepared with TURBOFRODO (http://www.afmb.univ-mrs.fr/-TURBO-), MOLSCRIPT[39] and RASTER3D[40].

31. Jones, L. R. Rapid preparation of canine cardiac sarcolemmal vesicles by sucrose flotation. *Methods Enzymol.* **157,** 85–91 (1988).
32. Skou, J. C. & Esmann, M. Preparation of membrane Na^+,K^+-ATPase from rectal glands of *Squalus acanthias*. *Methods Enzymol.* **156,** 43–46 (1988).
33. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
34. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
35. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
36. Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr. D* **55,** 247–255 (1999).
37. Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231,** 1049–1067 (1993).
38. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238,** 777–793 (1994).
39. Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24,** 946–950 (1991).
40. Merritt, E. A. & Bacon, D. J. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277,** 505–524 (1997).

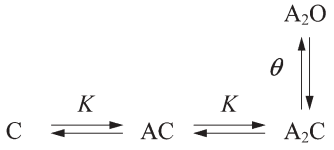
LETTERS

Detection and trapping of intermediate states priming nicotinic receptor channel opening

Nuriya Mukhtasimova[1]*, Won Yong Lee[1]*†, Hai-Long Wang[1] & Steven M. Sine[1,2]

In the course of synaptic transmission in the brain and periphery, acetylcholine receptors (AChRs) rapidly transduce a chemical signal into an electrical impulse. The speed of transduction is facilitated by rapid ACh association and dissociation, suggesting a binding site relatively non-selective for small cations. Selective transduction has been thought to originate from the ability of ACh, over that of other organic cations, to trigger the subsequent channel-opening step. However, transitions to and from the open state were shown to be similar for agonists with widely different efficacies[1–3]. By studying mutant AChRs, we show here that the ultimate closed-to-open transition is agonist-independent and preceded by two primed closed states; the first primed state elicits brief openings, whereas the second elicits long-lived openings. Long-lived openings and the associated primed state are detected in the absence and presence of an agonist, and exhibit the same kinetic signatures under both conditions. By covalently locking the agonist-binding sites in the bound conformation, we find that each site initiates a priming step. Thus, a change in binding-site conformation primes the AChR for channel opening in a process that enables selective activation by ACh while maximizing the speed and efficiency of the biological response.

Throughout the nervous system, moment-to-moment communication relies on the rapid on and off responses of synaptic receptors. Rapid switching is possible through a neurotransmitter-binding site freely accessible to solvent, enabling diffusion-limited binding of neurotransmitter, and modest stabilization of the bound complex, enabling quick release. Both adaptations seemingly oppose the ability of the binding site to capture the neurotransmitter from a mix of chemically similar molecules within the synapse. However, in light of the del Castillo-Katz model proposed fifty years ago[4], preferential activation by neurotransmitter over other organic cations could be encoded by a process now recognized as the channel-gating reaction, in which the fully occupied receptor channel reversibly opens and closes.



Modified to include two rather than one agonist-binding steps, this extended del Castillo-Katz model depicts binding of successive agonist molecules A to a receptor in the closed state C, generating an inactive complex A_2C, which isomerizes to the biologically active complex A_2O. Preferential activation by neurotransmitter could thus arise from its structurally encoded ability to drive the gating reaction in the forward direction, despite a relatively non-selective binding site.

The advent of patch-clamp recording enabled the registration of current pulses through single ion channels as brief as tens of

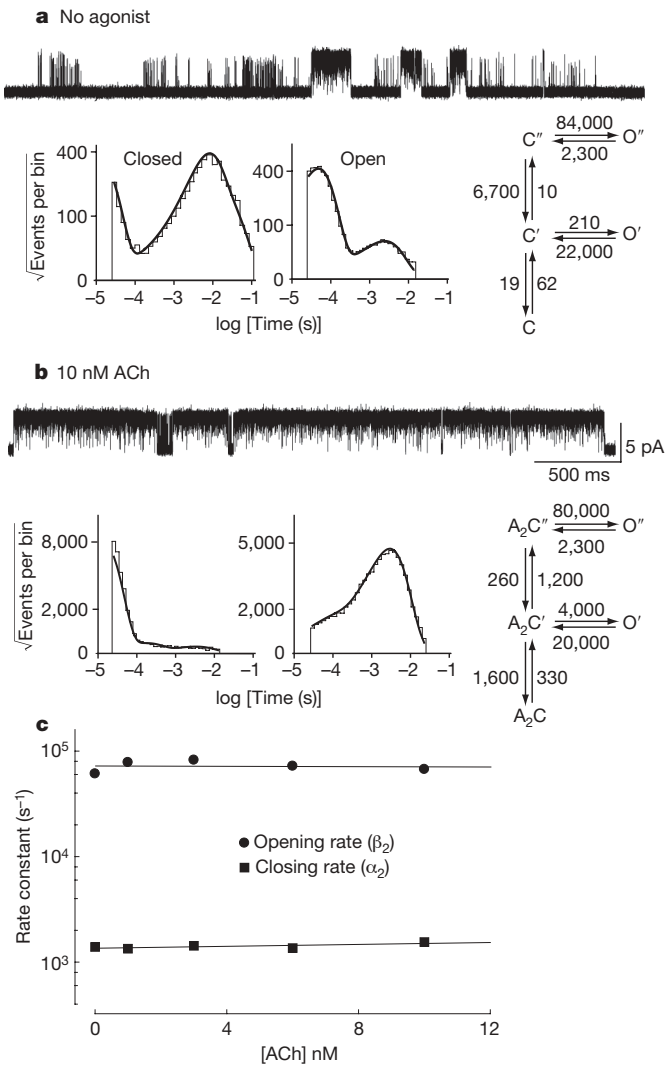


Figure 1 | Agonist-independent channel gating. a, Spontaneous single-channel currents through AChR containing Leu-to-Ser mutations at position 9′ of the second transmembrane domain (Leu9′Ser) of the β and δ subunits. Cell-attached configuration, membrane potential −70 mV; bandwidth 10 kHz; channel openings are upward deflections. Dwell-time histograms are shown with probability density functions obtained by fitting the scheme (right) to the dwell times. Rate constants are from Supplementary Table 1. **b**, Same as in **a** but with 10 nM ACh. **c**, Plot of fitted channel opening and closing rate constants against ACh concentration (Supplementary Table 1).

[1]Receptor Biology Laboratory, Department of Physiology and Biomedical Engineering, [2]Department of Neurology, Mayo Clinic College of Medicine, Rochester, Minnesota 55905, USA.
†Present address: Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA.
*These authors contributed equally to this work.

microseconds[5], allowing unprecedented testing of the binding–gating transduction mechanism. Channel openings were found to be interrupted by brief closed periods, and these were proposed to correspond to sojourns in the A_2C state within the extended del Castillo-Katz model[6]. By equating brief closed periods with sojourns in the A_2C state, kinetic modelling of current pulse sequences by methods pioneered by Colquhoun and Hawkes[7] yielded estimates of the rate at which fully occupied AChR channels opened[8,9]. The emerging single-channel opening rate approached the rate of rise of macroscopic current elicited by rapid application of agonist to AChR ensembles[10,11]. The overall data therefore supported the view that rapid onset of the post-synaptic response arises from low-affinity binding of ACh coupled directly to rapid channel opening.

However, a concurrent study found that brief closed periods that interrupted openings of single AChR channels were similar for agonists with widely different efficacies[1], contrary to the interpretation that brief closed periods corresponded to sojourns in the A_2C state. Building on kinetic studies of single glycine-activated receptors[12], a recent study of full and partial agonists provided evidence for a closed state called flipped, that followed agonist binding and preceded channel opening, but whose mean lifetime was similar for agonists with diverging efficacies[3]. Here, by studying mutant AChRs, we detect two distinct closed states, called primed, that follow agonist binding, couple tightly to channel opening and exhibit agonist-independent properties. Furthermore, we show that priming results from conformational changes at the two ACh-binding sites.

To look for agonist-independent transitions between closed and open states, we recorded single-channel currents through AChRs from adult skeletal muscle in the absence of agonist. Because wild-type AChRs rarely open spontaneously, we increased spontaneous opening by substituting Ser for a conserved Leu in the centre of the ion-conductive pore. In the absence of agonist, single AChR channels containing the Ser substitution in the β and δ subunits activate in long episodes of brief and long openings flanked by prolonged quiescent periods (Fig. 1a). Within such episodes, three closed and two open states are detected, indicating that the temporal sequence of single-channel current pulses arises from a minimum of five distinct states. Fitting a five-state model to the sequences of open and closed dwell times reveals that brief openings arise from a closed state with intermediate duration, whereas long openings arise from a closed state with brief duration (Fig. 1a and Supplementary Fig. 1). Notably, although agonist is not present, transitions from the brief-closed to the long-open state occur rapidly and with high probability.

Substituting Ser for the central Leu in other pairs of AChR subunits also increases spontaneous channel opening, which again appears as episodes of brief and long openings from a single receptor channel (Supplementary Fig. 2). A previous study documented an increase of spontaneous brief and long openings after substitution of Thr for a Ser approximately one turn of the pore helix from the Leu substituted here[13].

Application of ACh to our pore-mutant receptor increases long and decreases brief openings, with a tenfold change in ACh concentration increasing the fraction of long openings from 0.05 to greater than 0.90 (Fig. 1b). Fitting a five-state model to the sequences of ACh-evoked current pulses yields the surprising result that rates for entering and leaving the long-lived open state mimic those observed in the absence of agonist (Fig. 1c and Supplementary Table 1). Thus, transition from the major closed to the biologically relevant open state is independent of the presence of agonist.

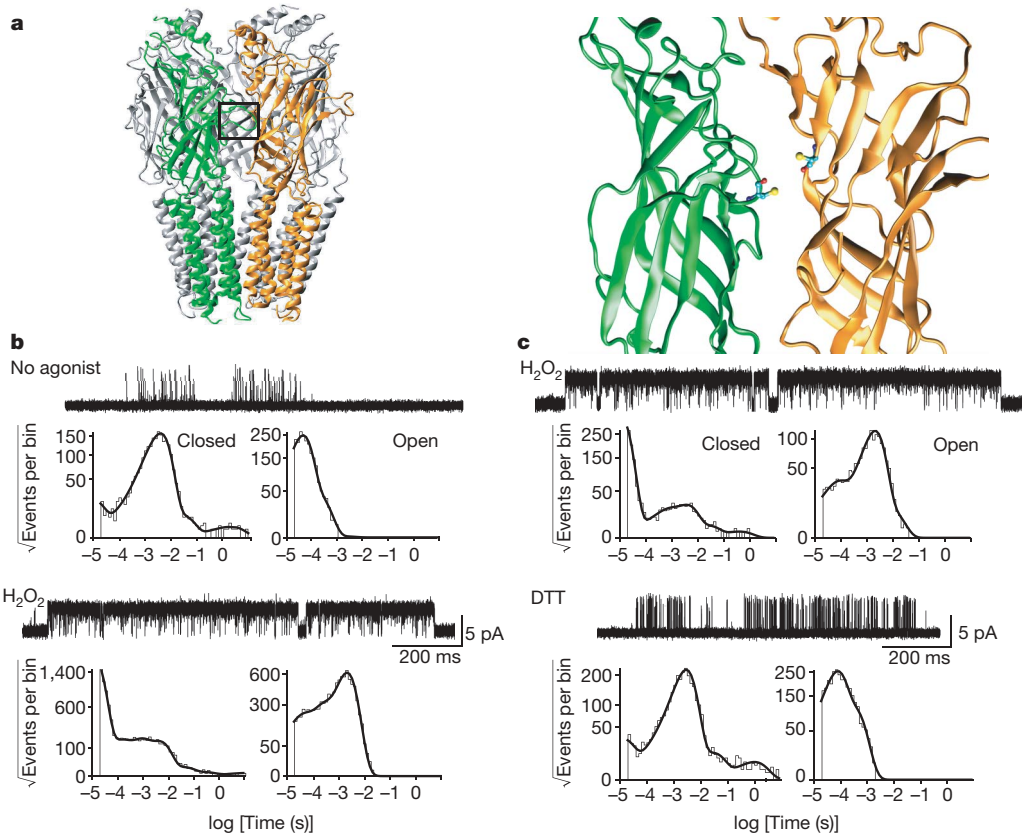


Figure 2 | Covalent priming of the AChR. **a**, *Torpedo* AChR (Protein Data Bank accession 2BG9); green, α subunit; orange, δ subunit. Boxed region (magnified to the right) indicates Cys substitutions at positions α192 and δ121. **b**, Upper trace: spontaneous currents through AChR containing Leu9′Ser mutations in the β and δ subunits, and Cys substitutions at both ACh-binding sites. Lower trace: spontaneous currents from the patch above after the application of H_2O_2. **c**, Upper trace: spontaneous currents through AChR with Leu9′Ser mutations and Cys substitutions at both ACh-binding sites after application of H_2O_2. Lower trace: spontaneous currents from the same patch but after application of dithiothreitol (DTT). In panels **b** and **c**, dwell-time histograms are fitted by sums of exponentials. Results are summarized in Supplementary Table 2.

To account for these observations, we propose that closed states that immediately precede channel opening correspond to AChRs with primed binding sites; priming of one site triggers brief openings, whereas priming of two sites triggers long openings. Comparison of crystal structures of the related ACh-binding protein, with and without bound agonist, shows that a hairpin structure flanking the binding site, called the C-loop, changes from an uncapped to a capped conformation after binding agonist[14,15]. Thus, we further propose that each priming step results from transition of a C-loop from the uncapped to the capped conformation.

To test this idea, we engineered a Cys residue at the tip of the C-loop of each binding site of our pore-mutant receptor (Fig. 2a), and engineered another Cys in each of the two juxtaposed subunits. We then monitored channel opening, in the absence of agonist, before and after applying an oxidizing reagent. Before oxidation, single receptor channels activate in episodes of predominantly brief openings (Fig. 2b). After oxidation, however, receptor channels activate in episodes of long openings in quick succession (Fig. 2b and Supplementary Table 2), suggesting that covalent reaction arrests the C-loops in the capped conformation, generating the doubly primed state that triggers long-lived channel openings. The functional consequences of oxidation are reversible; after applying the oxidizing reagent and generating long channel openings, application of a reducing reagent restored brief openings (Fig. 2c and Supplementary Table 2). Application of oxidizing or reducing reagents had minimal effect on pore-mutant AChRs with fewer than four Cys substitutions (Supplementary Table 3).

Receptors with the Cys pair at both binding sites, but without Ser substitutions in the pore, showed only rare spontaneous channel opening and no change after applying the oxidizing reagent (data not shown), suggesting that the uncapped conformation of the C-loop predominates, rendering the inter-Cys spacing too great for cross-linking. However with Ser substitutions in the pore, the capped conformation predominates, enabling covalent reaction. This retrograde communication between binding site and pore confirms the expectation that the two distal locations communicate in a bidirectional manner.

We further reasoned that the ability to prime should depend on the efficiency of bidirectional communication. To test this idea, we mutated key residues within the binding-pore linkage pathway in our pore-mutant AChR and recorded single channel currents in the absence of agonist. In the wild-type AChR, the mutation αY190F suppresses[16] whereas αP272A enhances[17] ACh-induced channel opening. When αY190F is engineered into the pore-mutant AChR, channel opening episodes consist solely of brief openings (Fig. 3a), suggesting that αY190F allows only a single priming step; application of ACh, however, triggers episodes of long openings in quick succession, suggesting agonist overcomes the attenuated priming caused by the mutation. When αP272A is engineered into the pore-mutant receptor, channel-opening episodes consist solely of long openings (Fig. 3b), suggesting that αP272A promotes priming of both binding sites, analogous to the action of agonist. As a control we tested the mutation εP121L, which suppresses AChR activation of the wild-type AChR but is not a component of the principal linkage pathway[18]. When εP121L is engineered into the pore-mutant receptor, channel opening episodes contain both brief and long openings (Fig. 3c), similar to the pore-mutant receptor without εP121L (Fig. 1a). Thus mutations that alter the linkage between binding and pore domains suppress or enhance priming and consequently alter channel opening.

To explain our collective findings, we propose a primed model to describe activation of the AChR (Fig. 4). The model consists of closed states arrayed in three columns, one for each degree of agonist occupancy (0–2), and three rows, one for each degree of priming (0–2). Primed states, C′ and C′′, give rise to channel opening, whereas unprimed states do not. Each element of the array represents a theoretically possible state, although only a subset of the states may be experimentally detectable. The left vertical plane of states depicts priming and channel

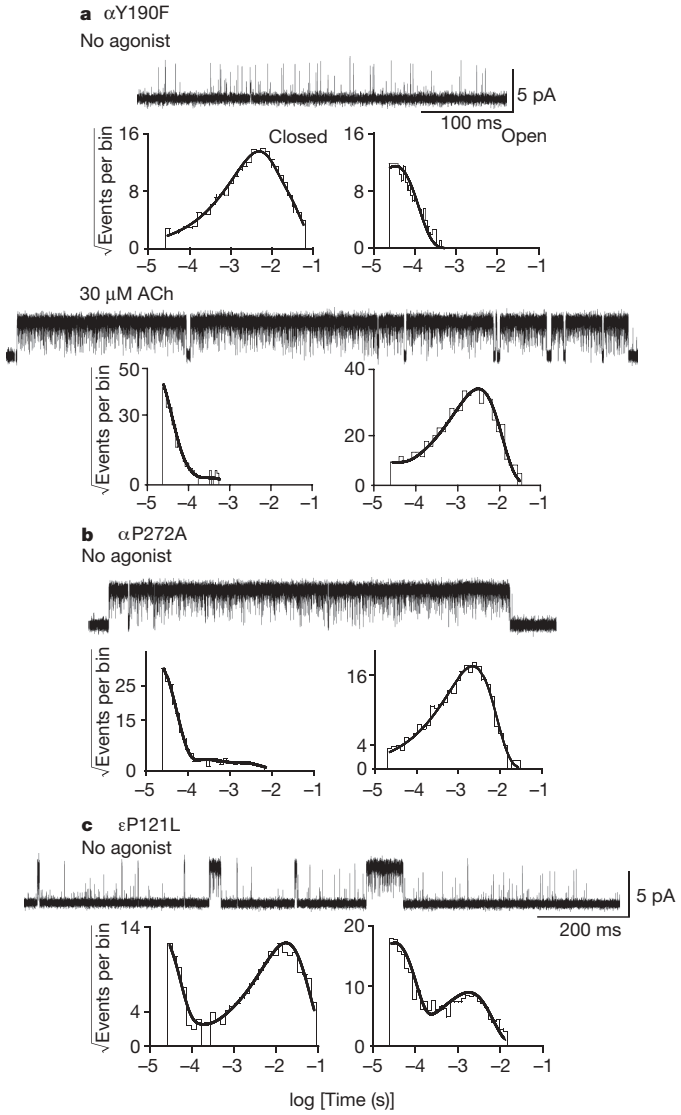


Figure 3 | Mutating residues linking binding and pore domains increases or decreases priming. **a**, Upper trace: spontaneous single channel currents through AChR containing Leu9′Ser mutations in the β and δ subunits and αY190F. Lower trace: ACh-evoked single-channel currents from a different patch containing the same mutant AChR. **b**, Spontaneous currents through AChR containing Leu9′Ser mutations in the β and δ subunits and αP272A. **c**, Spontaneous currents through AChR containing Leu9′Ser mutations in the β and δ subunits and εP121L. Dwell-time histograms are shown with fitted probability density functions (Supplementary Table 1).

opening in the absence of agonist, and in the wild-type AChR these transitions are rare. However, substituting hydrophilic residues in the pore unmasks states and reaction steps underlying unliganded channel gating; the singly primed state C′ elicits brief openings, whereas the doubly primed state C′′ elicits long openings. The right vertical plane depicts priming and channel opening with agonist bound to both binding sites, and in our pore-mutant AChR, transitions from the doubly primed to the long-open state predominate and mimic those that occur in the absence of an agonist.

As a further test, we fitted the primed model to sequences of single-channel closed and open dwell times obtained from the wild-type AChR activated by a wide range of ACh concentrations. Because not all states in the primed model are expected to occur with high probability, we fitted the most likely subset of the model to the data (Fig. 4). For comparison we fitted the extended del Castillo-Katz model to the data, as described previously[16,19]. Both models provide good descriptions of the dwell-time distributions (Supplementary Table 4 and Supplementary Fig. 3). However the computed log likelihood for the primed

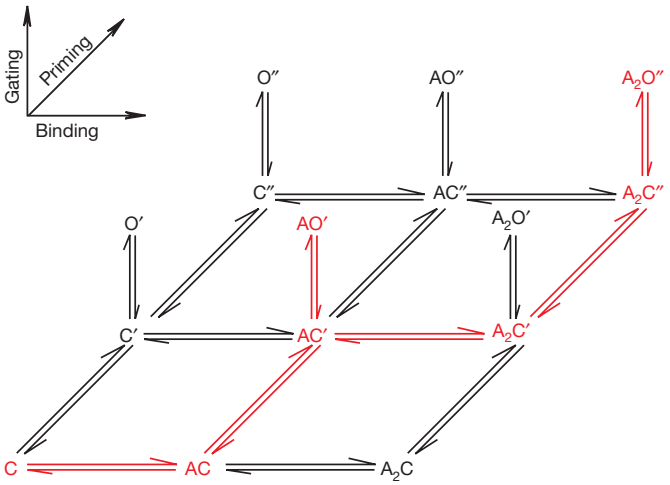


Figure 4 | Primed model of AChR activation. Agonist binding, priming and channel gating steps are indicated (inset). C, C′ and C′′ symbolize closed states, whereas O′ and O′′ symbolize open states. For the wild-type AChR in the absence of ACh, the C′ and C′′ states are negligible, indicating that the first step in the activation process generates AC, from which there are three possible paths towards A_2C''. Fitting the path bind–bind–prime–prime did not give well-defined rate constants, possibly owing to an inability to distinguish directly interconnected A_2C' and A_2C'' states. Fitting the path bind–prime–prime–bind also did not give well-defined rate constants, possibly because the second binding step would be to a primed site presumed to have reduced accessibility to small molecules. The remaining path in red was fitted to agonist-dependent dwell times from the wild-type AChR, yielding the rate constants in Supplementary Table 4.

model is significantly greater than that for the del Castillo-Katz model. The primed model yields greater ACh association and dissociation rate constants than the del Castillo-Katz model, suggesting that the binding site is more accessible to small molecules than previously recognized. In the primed model, the rate constant for generating the doubly primed closed state is similar to the rate constant for channel opening in the del Castillo-Katz model, which explains why the true channel opening step was obscured before.

More than two decades ago, biologically relevant channel openings of the AChR were found to be interrupted by agonist-independent brief closings[1,2]. Yet the functional significance of the interruptions remained elusive until a recent study of channel opening by partial agonists; the doubly occupied AChR was proposed to flip to a transient closed state, similar for all agonists, before the channel could open[3]. By studying mutant AChRs, we detect two transient closed states, called primed, tightly coupled to channel opening; the singly primed state has an intermediate duration and triggers brief openings, whereas the doubly primed state has a brief duration and triggers long-lived openings. Using disulphide trapping, we show that capping of each binding-site C-loop initiates priming, and that capping of both C-loops evokes long-lived openings. Our ability to unmask and modulate primed states is possible because of the bidirectional nature of communication between binding and pore domains. Priming of the AChR is thus a fundamental determinant of its biological activity; reduced priming by partial agonists would explain why they elicit a low maximal response but at the same time generate a stable open state interrupted by brief closings[1,3]. Priming seems to be an adaptation to endow the AChR with a rapid response, while maintaining preferential activation by agonist, thus preventing spurious responses to organic cations such as choline, a product of ACh hydrolysis. Furthermore, priming may represent a general adaptation by which chemically-mediated processes achieve both high speed and ligand specificity. Disruptions of the ability of a neurotransmitter to prime receptor channels for opening may underlie neurological diseases associated with the AChR and relatives in the Cys-loop superfamily.

METHODS SUMMARY

Construction of mutant AChR subunits, their expression in BOSC 23 cells, recordings of single channel currents and kinetic analyses of the currents are described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 October 2008; accepted 17 February 2009.
Published online 1 April 2009.

1. Sine, S. M. & Steinbach, J. H. Acetylcholine receptor activation by a site-selective ligand: nature of brief open and closed states in BC3H–1 cells. *J. Physiol. (Lond.)* **370,** 357–379 (1986).
2. Sine, S. M. & Steinbach, J. H. Activation of acetylcholine receptors on clonal mammalian BC3H–1 cells by low concentrations of agonist. *J. Physiol. (Lond.)* **373,** 129–162 (1986).
3. Lape, R., Colquhoun, D. & Silvilotti, L. On the nature of partial agonism in the nicotinic receptor superfamily. *Nature* **454,** 722–727 (2008).
4. del Castillo, J. & Katz, B. Interaction at end-plate receptors between different choline derivatives. *Proc. R. Soc. Lond. B* **146,** 369–381 (1957).
5. Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches. *Pflugers Arch.* **391,** 85–100 (1981).
6. Colquhoun, D. & Sakmann, B. Fluctuations in the microsecond time range of the current through single acetylcholine receptor ion channels. *Nature* **294,** 464–466 (1981).
7. Colquhoun, D. & Hawkes, A. On the stochastic properties of single ion channels. *Proc. R. Soc. Lond. B* **211,** 205–235 (1981).
8. Colquhoun, D. & Sakmann, B. Fast events in single-channel currents activated by acetylcholine and its analogues at the frog muscle end-plate. *J. Physiol. (Lond.)* **369,** 501–557 (1985).
9. Sine, S. M., Claudio, T. & Sigworth, F. J. Activation of *Torpedo* acetylcholine receptors expressed in mouse fibroblasts. Single channel current kinetics reveal distinct agonist binding affinities. *J. Gen. Physiol.* **96,** 395–437 (1990).
10. Liu, Y. & Dilger, J. P. Opening rate of acetylcholine receptor channels. *Biophys. J.* **60,** 424–432 (1991).
11. Maconochie, D. J. & Steinbach, J. H. The channel opening rate of adult- and fetal-type mouse muscle nicotinic receptors activated by acetylcholine. *J. Physiol. (Lond.)* **506,** 53–72 (1998).
12. Burzomato, V., Beato, M., Groot-Kormelink, P., Colquhoun, D. & Sivilotti, L. Single-channel behavior of heteromeric $\alpha1^{\beta}$ glycine receptors: an attempt to detect a conformational change before the channel opens. *J. Neurosci.* **24,** 10924–10940 (2004).
13. Grosman, C. & Auerbach, A. Kinetic, mechanistic and structural aspects of unliganded gating of nicotinic receptor channels: A single channel study of second transmembrane segment 12′ mutants. *J. Gen. Physiol.* **115,** 621–635 (2000).
14. Celie, P. H. *et al.* Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. *Neuron* **41,** 907–914 (2004).
15. Hansen, S. B. *et al.* Structures of *Aplysia* AChBP complexes with nicotinic agonists and antagonists reveal distinctive binding interfaces and conformations. *EMBO J.* **24,** 3635–3646 (2005).
16. Mukhtasimova, N., Free, C. & Sine, S. M. Initial coupling of binding to gating mediated by conserved residues in muscle nicotinic receptor. *J. Gen. Physiol.* **126,** 23–39 (2005).
17. Lee, W. Y., Free, C. & Sine, S. M. Nicotinic receptor inter-loop proline anchors β1-β2 and Cys-loops in coupling agonist binding to channel gating. *J. Gen. Physiol.* **132,** 265–278 (2008).
18. Ohno, K. *et al.* Congential myasthenic syndrome caused by decreased agonist binding affinity due to a mutation in the acetylcholine receptor ε subunit. *Neuron* **17,** 157–170 (1996).
19. Lee, W. Y. & Sine, S. M. Invariant aspartic acid in muscle nicotinic receptor contributes selectively to the kinetics of agonist binding. *J. Gen. Physiol.* **124,** 555–567 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Johnson for computer programming, and C. Free for technical contributions. This work was supported by National Institutes of Health grant NS031744 (S.M.S.).

Author Contributions N.M., W.Y.L. and H.L.W. conducted the experiments; N.M., W.Y.L. and S.M.S. analysed the data; S.M.S. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.M.S. (sine@mayo.edu).

METHODS

Construction of wild-type and mutant AChRs. Human α, β, δ and ε subunit complementary DNAs, subcloned in the CMV-based mammalian expression vector pRBG4 (ref. 20), were described previously[21]. Site-directed mutations were made using the QuickChange mutagenesis kit (Stratagene) and confirmed by sequencing the entire subunit cDNA. To generate a single free Cys at the tip of the C-loop of the α-subunit, Cys 193 was mutated to Ser, generating a free Cys at position 192.

Mammalian cell expression. All experiments used the BOSC 23 cell line[22], a variant of the 293 HEK cell line. Cells were maintained in DMEM containing fetal bovine serum (10% v/v) at 37 °C until they reached 50–70% confluence. Wild-type or mutant AChR cDNAs were transfected by calcium-phosphate precipitation using cDNA concentrations of 0.68 μg ml^{-1} for non-α-subunits and 1.36 μg ml^{-1} for the α subunit. Patch-clamp measurements were performed 1 or 2 days after transfection.

Patch-clamp recordings. To record single-channel currents, transfected cells were maintained in (mM): KCl 142, NaCl 5.4, $CaCl_2$ 1.8, $MgCl_2$ 1.7 and HEPES 10 (pH adjusted to 7.4). The same solution was used to fill patch pipettes. Acetylcholine (Sigma Chemical Co.) was kept as a 100 mM stock dissolved in bath solution and stored at –80°C until use. Glass micropipettes (type 7052, Garner Glass Co.), coated with Sylgard 184 (Dow Corning Co.), were heat-polished to yield resistances of 5–8 MΩ. Single-channel currents were recorded in the cell-attached configuration at 21 °C using an Axopatch 200B (Axon Instruments, Inc.) at a membrane potential of −70 mV. Data were collected from two to four different patches for each experimental condition; recordings were accepted for further analysis only when channel activity was low enough to clearly identify activation episodes from a single channel. The current signal was low-pass filtered at 50 kHz and recorded to hard disk at 200 kHz using the program Acquire (Bruxton Co.).

Single-channel kinetic analysis. The digitized current signal was filtered using a 10-kHz digital Gaussian filter[23], and channel events were detected by the half-amplitude threshold criterion using the program TAC (Bruxton Co.), using an imposed dead time of 10 μs. Precise determination of the dwell time at threshold was achieved by interpolating the digital signal and correcting the measured dwell time for the effects of the Gaussian filter[23]. Open and closed time histograms were fitted by the sum of exponentials using the program TACFit (Bruxton Co.). Openings corresponding to a single receptor channel were identified by assigning a critical closed time defined as the point of intersection of the longest closed time component, corresponding to closings between independent episodes of single channel openings, with the preceding briefer component, corresponding to closings between openings from a single channel. Kinetic analysis was performed using MIL software (QuB suite, State University of New York), which uses a maximum likelihood method, corrects for missed events and gives error estimates of the fitted rate constants[24]. An instrument dead time of 22 μs was uniformly applied to all recordings before fitting.

For kinetic analysis of wild-type AChRs, episodes of single channel currents containing at least five openings were analysed for open probability, mean open time and mean closed time, and episodes within two standard deviations of the mean were accepted for further analysis[25]. Kinetic analysis was performed by fitting models to data obtained across a range of ACh concentrations using MIL software. ACh concentrations ranged from 3 μM to 1 mM, with the concentrations spaced at half log unit intervals. Recordings from two patches for each ACh concentration were subjected to analyses, with an average number of events per patch of 4,208 (range 2,754–5,628); each kinetic model was fitted to data from all patches simultaneously. Data from different patches at the same ACh concentration were not pooled before fitting.

Disulphide trapping. After establishing a cell-attached gigaohm seal to BOSC 23 cells expressing Cys-substituted AChRs, spontaneous single-channel currents were recorded under control conditions. Freshly prepared H_2O_2 was then added to the bath solution, as described previously[26], to establish a final concentration of 8.8 mM, and a second recording was obtained after the change in current kinetics appeared complete (2–4 min).

To demonstrate reversal of the change in single-channel kinetics after oxidation, cells were treated with 8.8 mM H_2O_2 for 5 min, rinsed with normal bath solution, and a cell-attached gigaohm seal was established. After recording control spontaneous single channel currents, DTT was added to the bath solution to establish a final concentration of 0.02 mM, and a second recording was obtained after reversal of the current kinetics appeared complete (4–5 min)[26].

20. Lee, B. S., Gunn, R. B. & Kopito, R. R. Functional differences among nonerythroid anion exchangers expressed in a transfected human cell line. *J. Biol. Chem.* **266**, 11448–11454 (1991).
21. Ohno, K. *et al.* Congenital myasthenic syndrome caused by decreased agonist binding affinity due to a mutation in the acetylcholine receptor ε subunit. *Neuron* **17**, 157–170 (1996).
22. Pear, W. S., Nolan, G. P., Scott, M. L. & Baltimore, D. Production of high-titer helper-free retroviruses by transient transfection. *Proc. Natl Acad. Sci. USA* **90**, 8392–8396 (1993).
23. Colquhoun, D. & Sigworth, F. in *Single Channel Recording* (eds Sakmann, B. & Neher, E.) 191–264 (Plenum Publishing Corp., 1983).
24. Qin, F., Auerbach, A. & Sachs, F. Estimating single channel kinetic parameters from idealized patch clamp data containing missed events. *Biophys. J.* **70**, 264–280 (1996).
25. Wang, H.-L. *et al.* Mutation in the M1 domain of the acetylcholine receptor α subunit decreases the rate of agonist dissociation. *J. Gen. Physiol.* **109**, 757–766 (1997).
26. Mukhtasimova, N. & Sine, S. M. An intersubunit trigger of channel gating in the muscle nicotinic receptor. *J. Neurosci.* **27**, 4110–4119 (2007).

LETTERS

GlcNAcylation of a histone methyltransferase in retinoic-acid-induced granulopoiesis

Ryoji Fujiki[1,2], Toshihiro Chikanishi[1,2], Waka Hashiba[1], Hiroaki Ito[1], Ichiro Takada[1], Robert G. Roeder[3], Hirochika Kitagawa[1] & Shigeaki Kato[1,2]

The post-translational modifications of histone tails generate a 'histone code' that defines local and global chromatin states[1]. The resultant regulation of gene function is thought to govern cell fate, proliferation and differentiation[2]. Reversible histone modifications such as methylation are under mutual controls to organize chromosomal events[3,4]. Among the histone modifications, methylation of specific lysine and arginine residues seems to be critical for chromatin configuration and control of gene expression[5]. Methylation of histone H3 lysine 4 (H3K4) changes chromatin into a transcriptionally active state[6]. Reversible modification of proteins by β-*N*-acetylglucosamine (*O*-GlcNAc) in response to serum glucose levels regulates diverse cellular processes[7,8,9]. However, the epigenetic impact of protein GlcNAcylation is unknown. Here we report that nuclear GlcNAcylation of a histone lysine methyltransferase (HKMT), MLL5, by *O*-GlcNAc transferase facilitates retinoic-acid-induced granulopoiesis in human HL60 promyelocytes through methylation of H3K4. MLL5 is biochemically identified in a GlcNAcylation-dependent multi-subunit complex associating with nuclear retinoic acid receptor RARα (also known as RARA), serving as a mono- and di-methyl transferase to H3K4. GlcNAcylation at Thr 440 in the MLL5 SET domain evokes its H3K4 HKMT activity and co-activates RARα in target gene promoters. Increased nuclear GlcNAcylation by means of *O*-GlcNAc transferase potentiates retinoic-acid-induced HL60 granulopoiesis and restores the retinoic acid response in the retinoic-acid-resistant HL60-R2 cell line. Thus, nuclear MLL5 GlcNAcylation triggers cell lineage determination of HL60 through activation of its HKMT activity.

The currently known RAR co-regulators are unlikely to account for the pluripotent effects of retinoic acid (RA) in cell development[10]. Thus, we sought to identify new transcriptional regulators associated with RARα using biochemical analysis of human HL60 cells[11]. RARα-associated regulators were analysed in undifferentiated promyelocyte-like and differentiated granulocyte-like HL60 cells (Fig. 1a). A mass fingerprinting (matrix-assisted laser desorption/ionization–time of flight/mass spectrometry, MALDI–TOF/MS) analysis revealed 43 RARα interactants (Fig. 1b and Supplementary Table 1). MLL5 was chosen for further study because of its uncharacterized HKMT activity in multisubunit complexes[12–17]. Like the other MLL members, MLL5 harbours a SET domain, but its sequence bears little homology to those of the other MLL members (Supplementary Fig. 2).

The registered MLL5 gene consists of 27 exons encoding a ~200 kDa protein[14] (Supplementary Fig. 3a), although a much smaller MLL5 isoform (~75 kDa) was identified (Fig. 1b). By western blotting with two specific antibodies detecting the amino-terminal (anti-MLL5(N)) and carboxy-terminal (anti-MLL5(C)) regions of the MLL5 protein (Supplementary Fig. 3b), we found that the long form of the MLL5 protein (full-length MLL5, referred to here as MLL5Full) was expressed at much lower levels than the short form (MLL5) in undifferentiated HL60 cells (Fig. 1c) and in human tissues (Supplementary Fig. 4). By mapping a stop site for the short-form messenger RNA in front of exon 15 by rapid amplification of 3′-terminal cDNA

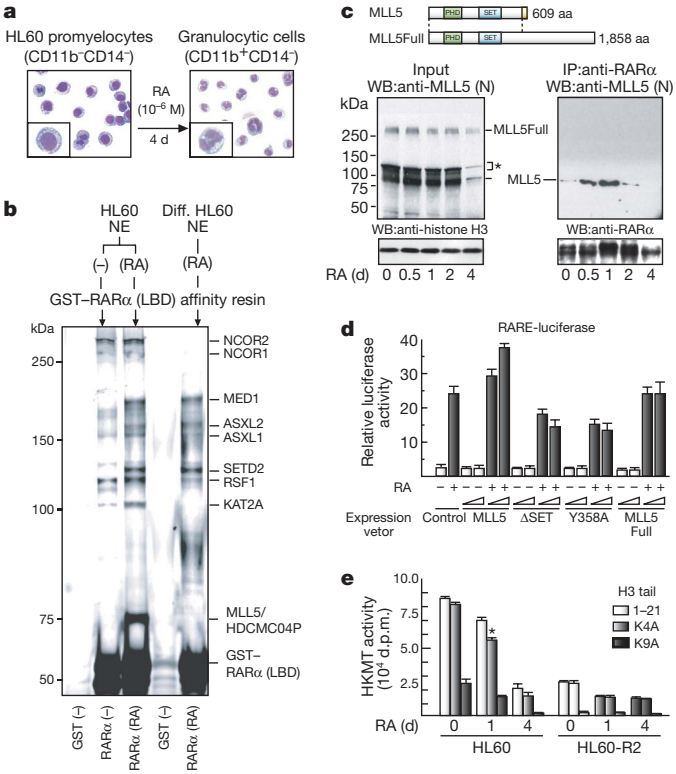


Figure 1 | MLL5 acts as a co-activator of RARα. **a**, Giemsa–May–Grunwald staining and cell-surface marker of HL60 cells. **b**, Outline (top) and silver-staining analysis (bottom) of the purification using GST-fused RARα (LBD) as bait. Diff. HL60, differentiated HL60 cells. **c**, Interaction between RARα and MLL5 during RA-induced differentiation. Protein structure of MLL5 (top). The anti-RARα immunoprecipitates (IP) from the RA-treated cells were subjected to western blotting (WB, bottom). The asterisk indicates a nonspecific band. **d**, Luciferase assay of MLL5 function in RAR-mediated transcription. RARE, RA response element. **e**, RARα-associating HKMT activities during RA-induced differentiation. The anti-RARα immunoprecipitants from the differentiating cells were used for *in vitro* HKMT assays with H3 tail peptides (1–21) and the indicated point-mutated peptides. $*P < 0.05$ versus the activity for the 1–21 peptide. Error bars, means and s.d. ($n = 3$). d.p.m., disintegrations per minute.

[1]Institute of Molecular and Cellular Biosciences, University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan. [2]ERATO, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchisi, Saitama 332-0012, Japan. [3]Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

ends (3′-RACE) (Supplementary Fig. 5), it was shown that the short isoform is the product of the registered MLL5 gene.

When overexpressed in a transient luciferase assay in undifferentiated HL60 cells, MLL5, but not MLL5Full, co-activated RARα (Fig. 1d), consistent with knockdown assays using short hairpin RNAs (shRNAs; Supplementary Fig. 6). When the SET domain (362−448 amino acid residues) was deleted (ΔSET) or inactivated (Y358A), RA-induced transactivation by means of RARα was inhibited (Fig. 1d and

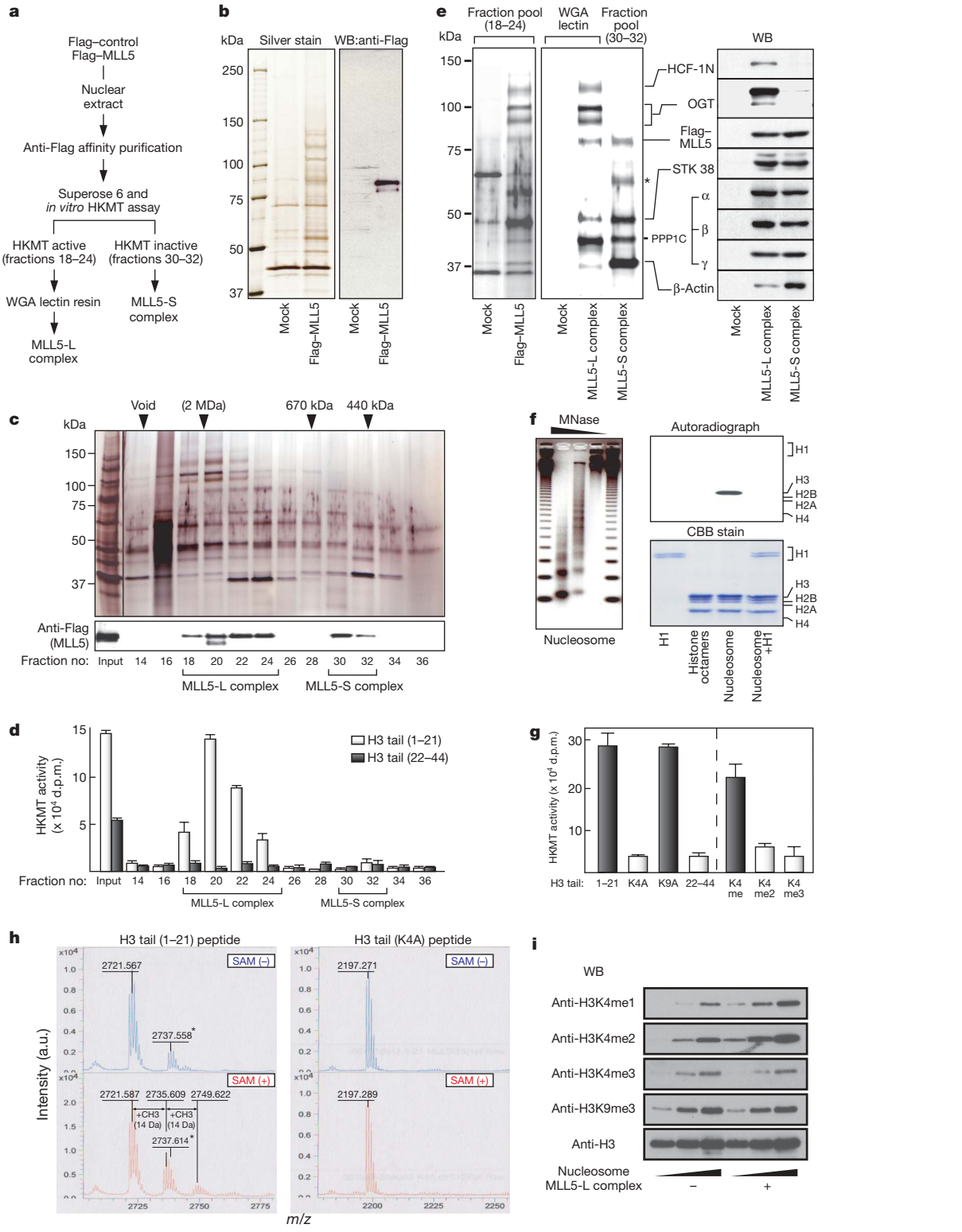


Figure 2 | Purification of the HKMT-active MLL5 complex. **a**, Outline of the purification. **b**, **c**, Silver staining and western blot analysis of the anti-Flag affinity purification for the Flag–MLL5 immunocomplex (**b**) and the gel-filtrated fractions (**c**). **d**, HKMT activities of the gel-filtrated fractions using the indicated tail peptides. **e**, Silver staining and western blot (WB) analysis of the MLL5-L and the MLL5-S complexes. The asterisk indicates a nonspecific band. **f**, **g**, Substrate preference of the MLL5-L complex. *In vitro* HKMT assay using the native histone substrates (**f**) and H3 tail peptides (**g**). *In vitro* reconstituted nucleosomes were analysed by micrococcal nuclease assay (MNase assay; **f**). CBB coomassie brilliant blue. **h**, **i**, Mass-spectrometric analysis of the H3 tail (1–21 and K4A) peptides (**h**) and western blot analysis of the nucleosomes (**i**) methylated by the MLL5-L complex. The asterisks indicate nonspecific peaks (**h**). Error bars, means and s.d. ($n = 3$). SAM, *S*-adenosyl-L-methionine.

Supplementary Fig. 7). In RARα immunoprecipitates from undifferentiated HL60 cells, significant H3K4 HKMT activity was detected for several histone H3 substrates (H3K4, H3K9, H3K27 and H3K36; Fig. 1e and Supplementary Fig. 8). However, unexpectedly, recombinant MLL5 proteins bacterially prepared were unable to elicit H3K4 HKMT activity *in vitro* (data not shown).

H3K4 HKMT activity was at negligible levels in the HL60-R2 cell line (Fig. 1e), even though MLL5 was expressed at a comparable level (Supplementary Fig. 9). This line is resistant to RA-induced cellular differentiation through an unknown mechanism that does not involve an RARα gene mutation[18].

An MLL5 complex was isolated from a newly established HL60 stable transformant expressing Flag-tagged MLL5 (Fig. 2a and Supplementary Fig. 10). Corresponding to anti-Flag−MLL5 immunocomplexes (Fig. 2b), Fig. 2c shows two peaks with apparent masses of ∼1.5 MDa (fractions 18−24, MLL5-L) and ∼600 kDa (fractions 30−32, MLL5-S). The MLL5-L fraction contained HKMT activity, whereas activity was marginal in the MLL5-S fractions (Fig. 2d).

Two different complexes shared the following common components: serine/threonine kinase 38 (STK38), protein phosphatase 1 catalytic subunits (α, β and γ isoforms; PPP1Cα, β and γ) and β-actin (Fig. 2e, left). Of note, the HKMT-active MLL5-L complex also contained the host cell factor-1 N-terminal subunit (HCF-1N, also known as HCFC1) and *O*-GlcNAc transferase (OGT). Because OGT efficiently binds to wheat germ agglutinin (WGA)[19], we verified the presence of OGT after purification on a WGA column. The eluates then represented the stoichiometric composition of the MLL5-L complex (Fig. 2e, middle), which was further confirmed by western blotting (Fig. 2e, right).

The purified MLL5-L complex, but not MLL5-S (data not shown), methylated nucleosomal H3 in a reconstituted chromatin template (Fig. 2f). We assessed the substrate specificity of the purified MLL5-L complex with several peptides and an artificial H3 N-terminal peptide (1−21), in which Lys 4 was replaced with Ala (K4A). H3K4 was determined to be the MLL5 HKMT target residue (Fig. 2g). Mono-, but not di- nor tri-, methylated H3K4 peptide (1−21) was a good substrate. Methylations were further confirmed by MALDI–TOF/MS (Fig. 2h and Supplementary Fig. 11) and western blotting (Fig. 2i).

On the basis of the inclusion of OGT in the large, active complex, we hypothesized that OGT activates the complex through GlcNAc transfer. GlcNAcylation of the purified complex was confirmed by western blotting with anti-GlcNAc monoclonal antibody (clone number CTD110.6) that specifically recognizes GlcNAcylated Ser/Thr residues (Fig. 3a). Recombinant OGT protein could GlcNAcylate recombinant MLL5 protein *in vitro* (Supplementary Fig. 12) and induce H3K4 HKMT activity (Supplementary Fig. 13).

Addition of the *O*-glycosyl donor UDP–GlcNAc potentiated HKMT activity of the MLL5-L complex, whereas a decoy substrate of OGT, UDP–GalNAc, lacked such an effect (Fig. 3b). The purified MLL5 complex completely lost HKMT activity (Fig. 3b) on conversion into

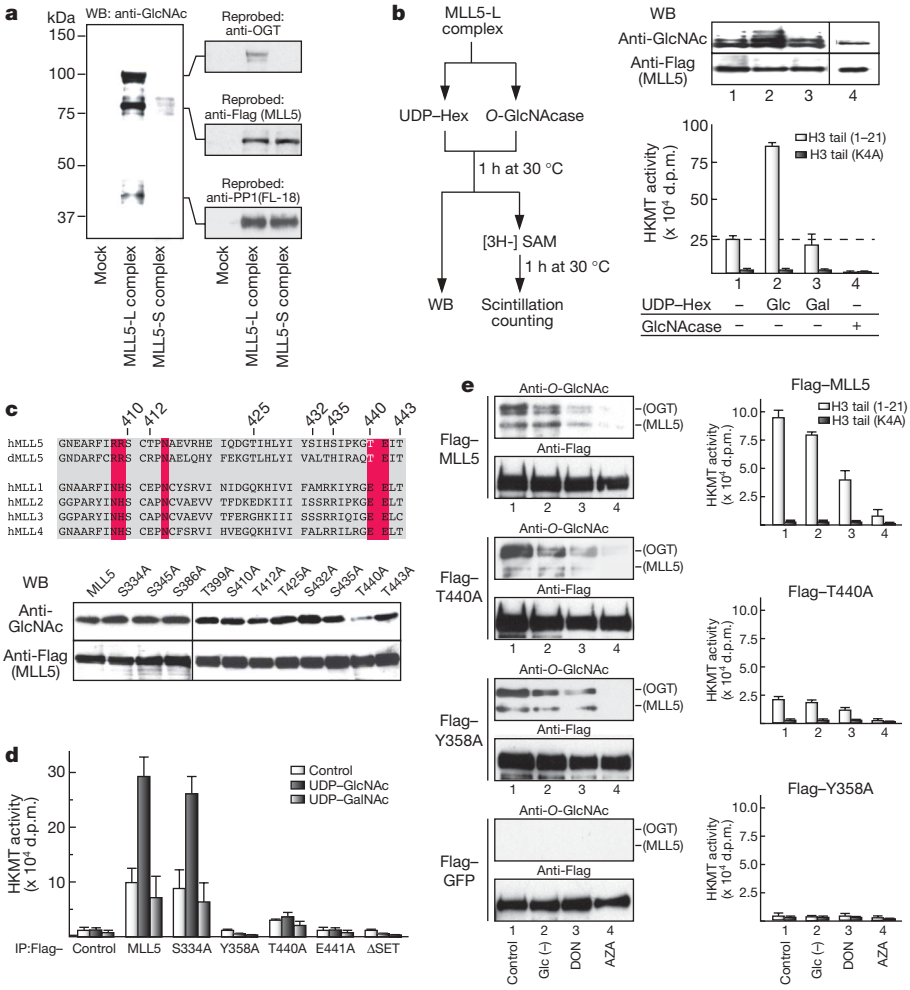


Figure 3 | MLL5 is a GlcNAcylation-dependent HKMT. **a**, Western blot (WB) analysis for GlcNAcylation of the MLL5-L and the MLL5-S complexes. The anti-GlcNAc bands were re-probed with the indicated antibodies. **b**, Effect of *in vitro* GlcNAcylation of the MLL5-L complex on its HKMT activity. The experimental procedure is summarized (left). Gal, UDP–GalNAc; Glc, UDP–GlcNAc. **c**, Mapping of GlcNAcylation sites of MLL5. Sequence alignment of MLL-family SET domains (top). The residues required for SAM binding are shaded in red. Western blot analysis for GlcNAcylation of the serial point-mutated MLL5s (indicated Ser/Thr to Ala, bottom). **d**, **e**, GlcNAc-modified (**d**) or -depleted (**e**) MLL5 mutants were subjected to the HKMT assay. GlcNAc levels were analysed by western blot (**e**). Error bars, means and s.d. ($n = 3$).

a smaller complex (Supplementary Fig. 14) when the GlcNAcylation moieties of the complex were removed by hexosaminidase (*O*-GlcNAcase) treatment. Thus, GlcNAcylation of OGT and MLL5 appeared necessary to form a large and enzymatically active form of the MLL5 complex.

To define GlcNAcylation site(s) of MLL5, we generated deletion mutants (Supplementary Fig. 15) and individually replaced all 11 Ser and Thr residues in the SET domain with Ala (Fig. 3c). Western blotting identified Thr 440 as a major GlcNAcylation site (Fig. 3c, bottom). This Thr residue is conserved in a putative fly MLL5 homologue (CG9007), but is not present in other MLL family members (Fig. 3c, top). Immunoprecipitates of the MLL5 T440A mutant were enzymatically defective like other catalytically dead SET-domain mutants (Y358A and E441A mutants, Fig. 3d) without overt alteration of protein structure (Supplementary Figs 16 and 17). Potentiation of HKMT activity by UDP–GlcNAc was not seen in similarly prepared immunoprecipitates of the other MLL family members (Supplementary Fig. 18).

MLL5 GlcNAcylation and its HKMT activity depend on the presence of sufficient levels of nuclear UDP–GlcNAc[7] (Fig. 3e). UDP–GlcNAc is derived from extracellular glucose through the cellular hexosamine biosynthesis pathway (HBP)[7]. Cells were treated with either high-glucose (10–30 mM) media or a GlcNAcase inhibitor (PUGNAc, ~150 μM) to fully induce nuclear protein GlcNAcylation. Both treatments effectively potentiated RA-induced differentiation of HL60 into granulocytes (Fig. 4a), but not into other cell types (Supplementary Fig. 19). Suppression of HBP by specific inhibitors (6-diazo-5-oxo-L-norleucine, DON, and azaserine, AZA), which reduced nuclear UDP–GlcNAc level as well as the GlcNAcylation level of MLL5 (Supplementary Fig. 20), attenuated the RA effect (Fig. 4a). This was restored by glucosamine treatment (Supplementary Fig. 21).

Endogenous HKMT activity of MLL5 was much lower in HL60-R2 cells than in HL60 cells (Supplementary Fig. 22). However, treatment with PUGNAc restored responses to RA (Fig. 4b and Supplementary Fig. 23) and RA-induced methylation of histone H3K4 (Fig. 4c and Supplementary Fig. 24). MLL5 was less GlcNAcylated in HL60-R2 cells than in HL60 cells (Fig. 4d), and untreated HL60-R2 cells contained very low levels of intracellular glucose and high *O*-GlcNAcase activity in comparison to HL60 cells (Supplementary Fig. 25), indicating nuclear hypo-GlcNAcylation in HL60-R2 cells.

We monitored RA-induced gene expression of C/EBPε (also known as CEBPE; a major granulopoietic regulator in HL60 cells[20]) to test the impact of MLL5. C/EBPε gene induction by RA was potentiated by PUGNAc (Fig. 4e, left). Knockdown of MLL5 or OGT by shRNA abrogated RA-induced gene induction. Overexpression of wild-type MLL5, but not the T440A mutant, potentiated the RA effect in HL60 cells, and PUGNAc further enhanced the MLL5 function (Fig. 4e, right). Chromatin immunoprecipitation (ChIP) analysis showed

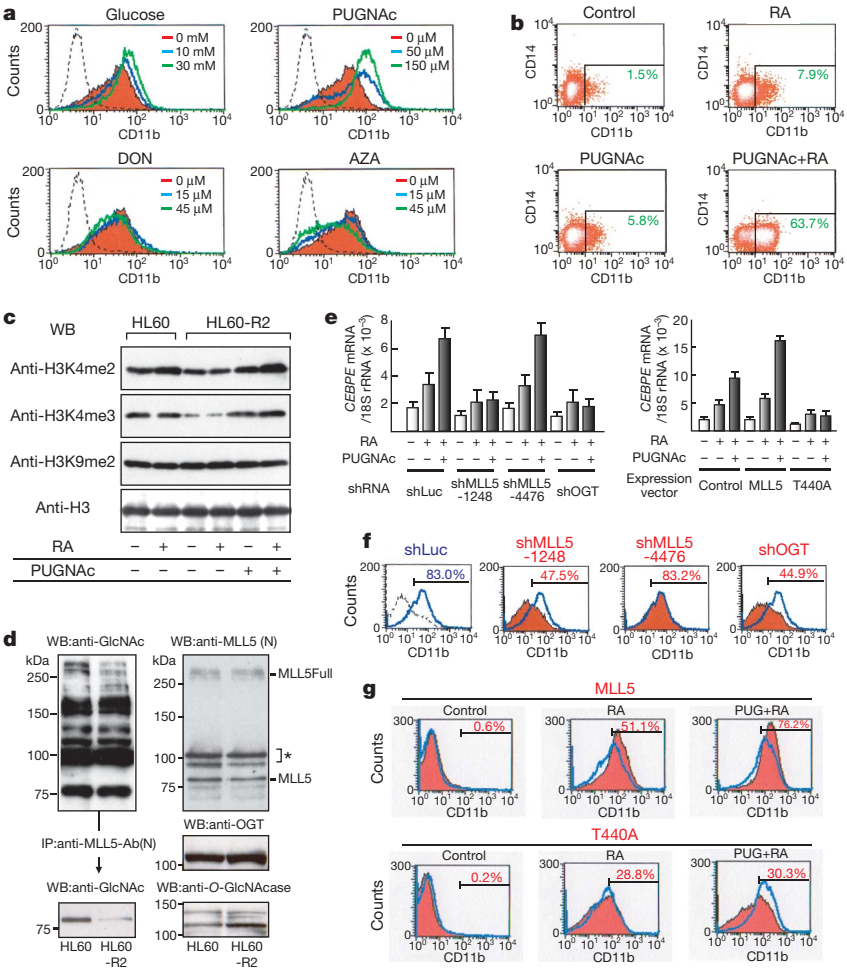


Figure 4 | GlcNAcylation of MLL5 facilitates RA-induced granulopoiesis. **a**, **b**, Effect of cellular GlcNAcylation on RA-induced granulopoiesis. The HL60 cells (**a**) or the HL60-R2 cells (**b**), exposed to the indicated reagents, were analysed by flow cytometry. The dashed line shows RA-untreated control (**a**). **c**, Western blot (WB) analysis for histone H3K4 methylation in the cells treated with RA, PUGNAc or both. **d**, Western blot analysis for GlcNAcylation of MLL5 and OGT in HL60 and HL60-R2 cells. **e**, Quantitative polymerase chain reaction (qPCR) analysis of *CEBPE* expression in the cells retrovirally transduced with the indicated shRNAs or expression vectors (using 18S rRNA as internal control). **f**, **g**, The roles of MLL5 and OGT on RA-induced differentiation. RA-induced differentiation of the HL60 cells, virally expressed with the indicated shRNAs (**f**) or the indicated constructs (**g**), were analysed by flow cytometry. shLuc (**f**) or Flag (**g**) control is overlaid (blue-open). Error bars, means and s.d. ($n = 3$).

RA-induced recruitment of RARα and MLL5 complex components to the C/EBPε gene promoter and H3K4 methylation (Supplementary Fig. 26). Similar results were seen in another RARα target gene, *RARB* isoform 2 (*RARB2*) (refs 10 and 11; Supplementary Fig. 27).

The physiological importance of MLL5 in RA-induced granulopoiesis was tested by MLL5 knockdown using shRNA (Supplementary Fig. 28). In HL60 cells expressing shMLL5-1248 (targeting MLL5 and MLL5Full) or shOGT (based on ZsGreen expression), the capacity of RA to induce granulopoiesis was impaired by 47.5% or 44.9%, respectively (Fig. 4f). It was not inhibited by shRNAs targeting other MLL members (Supplementary Fig. 29).

An enhanced granulopoietic response to RA was detected in HL60 cells overexpressing MLL5 via retroviral vector (Fig. 4g, upper). Under the same experimental conditions, RA-induced granulopoiesis was attenuated in cells expressing either the MLL5 T440A mutant (Fig. 4g, bottom) or the catalytically dead SET domain mutants (Supplementary Fig. 30), presumably by acting as 'a dominant-negative mutant' for endogenous GlcNAcylated MLL5.

Here we show that nuclear protein GlcNAcylation facilitates RA-induced differentiation by directly activating MLL5 to di-methylate histone H3K4 (Supplementary Fig. 1). OGT is an indispensable component of the MLL5-L complex required for the GlcNAcylation of MLL5. Recently, two groups reported that, because UDP−GlcNAc levels are controlled through the HBP in response to serum levels of circulating glucose[21,22], reversible GlcNAcylation of cytosolic proteins may represent a sensory system for glucose homeostasis. Intracellular transport of OGT from the nucleus to the plasma membrane induced by insulin/phosphoinositide signalling was shown[21]. However, unlike an insulin signal, RA does not induce re-localization of nuclear OGT in HL60 cells (Supplemental Fig. 31). Together with these previous findings[8,9,21,22], it appears that protein GlcNAcylation by OGT is a common post-translational modification of cytosolic and nuclear proteins. Thus, our findings imply that nuclear protein GlcNAcylation is a vital fine-tuning effector in epigenetics.

METHODS SUMMARY

MLL5 and *MLL5Full* cDNAs were isolated from HL60 cells. Both HL60 and HL60-R2 cell lines were maintained in RPMI medium (10% FBS). For HL60 cell granulopoiesis, the cells were treated for 4 days with RA (1 μM). HL60 cells were infected with the retrovirus under centrifugation at 1,000*g* for 1 h at room temperature (25 °C) in the presence of polybrene (30 μg ml^{-1}). For assessment of RA-induced differentiation of the infected cells, the cells expressing enhanced green fluorescent protein (eGFP) or ZsGreen, along with MLL5 constructs or MLL5-targeted shRNAs, were analysed by flow cytometry with allophycocyanin (APC)-conjugated anti-CD11b (also known as ITGAM). Biochemical purification using recombinant GST-fused ligand-binding domain of human RARα (GST−RARα LBD; 153−462 amino acids) bound to glutathione sepharose was performed as previously described[23,24]. Nuclear extracts (0.5 g) prepared from RA-treated or -untreated HL60 cells were incubated with 5 μg of human GST−RARα LBD with or without RA (1 μM). The tryptic polypeptides of the individually excised bands were analysed by MALDI–TOF/MS and identified from the peptide mass fingerprints. For purification of the MLL5 complex, HL60 cells stably expressing Flag-tagged MLL5 along with the eGFP reporter were concentrated by fluorescence-activated cell sorting (FACS). Using nuclear proteins (0.5 g) derived from the stably transfected cells, the MLL5-L complex was isolated by monitoring HKMT activity with the following three purification steps: anti-Flag affinity, gel filtration chromatography with Superose 6 column (GE Healthcare) and WGA lectin agarose. The *in vitro* HKMT and OGT assays were performed as previously reported[25–27]. Recombinant Flag–MLL5 and 6 × His-OGT proteins were prepared by the Bac-to-Bac baculovirus expression system (Invitrogen).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 14 November 2008; accepted 10 March 2009.
Published online 19 April 2009.

1. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).
2. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128,** 669–681 (2007).
3. Sarma, K. & Reinberg, D. Histone variants meet their match. *Nature Rev. Mol. Cell Biol.* **6,** 139–149 (2005).
4. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
5. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15,** 2343–2360 (2001).
6. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
7. Hart, G. W., Housley, M. P. & Slawson, C. Cycling of *O*-linked β-*N*-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **446,** 1017–1022 (2007).
8. Jackson, S. P. & Tjian, R. *O*-glycosylation of eukaryotic transcription factors: implications for mechanisms of transcriptional regulation. *Cell* **55,** 125–133 (1988).
9. Kelly, W. G. & Hart, G. W. Glycosylation of chromosomal proteins: localization of *O*-linked *N*-acetylglucosamine in *Drosophila* chromatin. *Cell* **57,** 243–251 (1989).
10. Chambon, P. A decade of molecular biology of retinoic acid receptors. *FASEB J.* **10,** 940–954 (1996).
11. Collins, S. J. The role of retinoids and retinoic acid receptors in normal hematopoiesis. *Leukemia* **16,** 1896–1905 (2002).
12. Nakamura, T. *et al.* ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Mol. Cell* **10,** 1119–1128 (2002).
13. Dou, Y. *et al.* Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF. *Cell* **121,** 873–885 (2005).
14. Emerling, B. M. *et al. MLL5*, a homolog of *Drosophila trithorax* located within a segment of chromosome band 7q22 implicated in myeloid leukemia. *Oncogene* **21,** 4849–4854 (2002).
15. Hughes, C. M. *et al.* Menin associates with a trithorax family histone methyltransferase complex and with the hoxc8 locus. *Mol. Cell* **13,** 587–597 (2004).
16. Lee, M. G. *et al.* Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science* **318,** 447–450 (2007).
17. Cho, Y. W. *et al.* PTIP associates with MLL3- and MLL4-containing histone H3 lysine 4 methyltransferase complex. *J. Biol. Chem.* **282,** 20395–20406 (2007).
18. Mori, J. *et al.* Characterization of two novel retinoic acid-resistant cell lines derived from HL-60 cells following long-term culture with all-trans-retinoic acid. *Jpn. J. Cancer Res.* **90,** 660–668 (1999).
19. Wysocka, J. *et al.* Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3–K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev.* **17,** 896–911 (2003).
20. Si, J., Mueller, L. & Collins, S. J. CaMKII regulates retinoic acid receptor transcriptional activity and the differentiation of myeloid leukemia cells. *J. Clin. Invest.* **117,** 1412–1421 (2007).
21. Yang, X. *et al.* Phosphoinositide signalling links *O*-GlcNAc transferase to insulin resistance. *Nature* **451,** 964–969 (2008).
22. Dentin, R. *et al.* Hepatic glucose sensing via the CREB coactivator CRTC2. *Science* **319,** 1402–1405 (2008).
23. Kitagawa, H. *et al.* The chromatin-remodeling complex WINAC targets a nuclear receptor to promoters and is impaired in Williams syndrome. *Cell* **113,** 905–917 (2003).
24. Ohtake, F. *et al.* Dioxin receptor is a ligand-dependent E3 ubiquitin ligase. *Nature* **446,** 562–566 (2007).
25. Takada, I. *et al.* A histone lysine methyltransferase activated by non-canonical Wnt signalling suppresses PPAR-gamma transactivation. *Nature Cell Biol.* **9,** 1273–1285 (2007).
26. Fujiki, R. *et al.* Ligand-induced transrepression by VDR through association of WSTF with acetylated histones. *EMBO J.* **24,** 3881–3894 (2005).
27. Wells, L. *et al. O*-GlcNAc transferase is in a functional complex with protein phosphatase 1 catalytic subunits. *J. Biol. Chem.* **279,** 38466–38470 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Miyajima and S. Saito for cell lines, P. Chambon, T. Kitamura and S. Suzuki for experimental materials, H. Akaishi for technical support, and M. Yamaki for manuscript preparation. This work was supported in part by priority areas from the Ministry of Education, Culture, Sports, Science and Technology (to H.K. and S.K.).

Author Contributions S.K. planned the project and analysed the experiments together with R.F., R.G.R. and H.K. R.F., T.C., W.H., H.I. and I.T. conducted the experiments. The manuscript was written by S.K. and R.F., and all authors commented on it.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.K. (uskato@mail.ecc.u-tokyo.ac.jp).

METHODS

Plasmids and retroviruses. Full-length human *MLL5* and *MLL5Full* gene transcripts were amplified from mRNA from HL60 cells. cDNAs of *MLL5*, *MLL5Full* and their deletion mutants, which are fused with Flag tag sequence at their 5′ terminals, were subcloned into pMXIG (provided by T. Kitamura) or pcDNA3 (Invitrogen). A series of point mutations in the vectors encoding *MLL5* cDNA was generated with the QuikChange II site-directed mutagenesis kit (Stratagene). shRNA sequences targeting human MLL5-742 (5′-GGATCTAGG AAGTCATCAAGA-3′), human MLL5-1248 (5′-GGTGAGGCATGAAATTCA AGA-3′), human MLL5Full-4476 (5′-GCAACACTCTGTAGCACATGT-3′), human OGT (5′-GCACATAGCAATCTGGCTTCC-3′) and *Renilla* Luc (5′-TG CGTTGCTAGTACCAAC-3′, as a non-targeting control) were inserted into the pSIREN-RetroQ-ZsGreen vector (Clontech). For retroviral production, the constructed pMXIG and pSIREN-RetroQ-ZsGreen vectors were transfected into PLAT-A cells (provided by T. Kitamura) with Lipofectamine 2000 reagent (Invitrogen). The culture medium was replaced with fresh DMEM containing 10% FBS 24 h after the transfection, and the cells cultured further for an additional 24 h. The culture supernatant containing the virus was then used for infection.

Cell culture, retroviral infection, differentiation and Giemsa–May–Grunwald staining. HL60 cells were maintained in RPMI medium supplemented with 10% FBS. For retroviral infection, 10^6 cells were suspended in 2 ml of retroviral cocktail (1 ml of the prepared retroviral solution plus 1 ml of RPMI with 10% FBS and 60 μg ml^{-1} polybrene) and then centrifuged at 1,000*g* for 1 h at room temperature. For granulocytic or monocytic differentiation of HL60 cells, 10^6 cells in the exponential growth phase were exposed to 1 μM RA (Sigma) or 1 μM VD (1α,25(OH)$_2$ vitamin D$_3$, WAKO). The differentiated cells were cytospun on a glass slide and stained with Giemsa solution (Sigma) and May–Grunewald solution (Sigma) according to the manufacturer's instructions.

Flow cytometric analysis and sorting. For flow cytometric evaluation of RA- or VD-induced differentiation, ligand-stimulated cells (10^6) were incubated with 10 μg ml^{-1} human immunoglobulin G (Sigma) and then immunostained with Alexa Fluor 488- or APC-conjugated anti-human CD11b (BD biosciences), and APC-conjugated anti-human CD14 (BD biosciences). In the presence of 10 μg ml^{-1} propidium iodide (Sigma), 2×10^4 cells labelled with the antibodies were analysed with the FACS Caliber system (BD biosciences). A FACS Vantage (BD biosciences) sorter was used to isolate the retroviral-transduced, eGFP-positive cells.

Electroporation and luciferase reporter assay. For the luciferase assays, electroporation was performed by Gene pulser MXcell system (Bio-Rad) at 300 V, 2,000 μF, 1,000 Ω and 10 ms (square-wave pulse). Cells (5×10^6) at the early log phase of growth were electroporated with 25 μg of the indicated plasmids (including 10 μg pGL3 reporters (Promega), 10 μg pRL-null reporters (Promega), 5 μg expression vectors encoding the MLL5 and OGT constructs) in RPMI medium. The electroporated cells were further cultured in the presence or absence of ligand for 48 h. Luciferase activities expressed in the cells were measured with the Dual Luciferase Assay System (Promega).

Immunoprecipitation and immunoblotting. For immunoprecipitation and immunoblotting, whole-cell lysates were prepared with TNE (20 mM Tris-HCl, 137 mM NaCl, 10% (v/v) glycerol, 1% (v/v) NP-40, 2 mM EDTA, pH 7.9). In the presence or absence of 1 μM RA, lysates (1 ml) were immunoprecipitated with anti-RARα (Santa Cruz) or anti-Flag M2 (Sigma). Immunoblotting was performed with anti-MLL5(N) (Operon Biotechnology), anti-MLL5(C) (MBL), anti-OGT (Sigma), anti-*O*-GlcNAc (CTD110.6, Covance), anti-HCF-1N (Bethyl), anti-STK38 (Abnova), anti-PPP1Cα (Upstate), anti-PPP1Cβ (Chemicon), anti-PPP1Cγ (Sigma), anti-β-actin (Santa Cruz) and anti-Flag (Sigma). For the generation of an anti-MLL5(N) antibody, three synthetic peptides, NH$_2$-Cys+RLGNDKKEMNKS-COOH, NH$_2$-CEGTTNKMKSPETKQ-COOH and NH$_2$-Cys+QEPDFIDDIEEKTP-COOH, individually conjugated to KLH were used as immunogens. To generate MLL5(C) antibody, three KLH-conjugated synthetic peptides, NH$_2$-Cys+LMEDPDPENPEPTTNEC-COOH, NH$_2$-CSEKNEKTGKPSDGLSER-COOH and NH$_2$-CAQVPPTFQNNYHGSGWH-COOH, were used as immunogens. For immunoblotting of histone methylation, total cell lysates were prepared by sonication in RIPA buffer (50 mM Tris-HCl, 150 mM NaCl, 1% (v/v) NP-40, 0.5% sodium deoxycholate (w/v), 0.1% SDS (w/v), pH 7.9).

ChIP assay. For ChIP analysis, the experimental procedure was basically performed as previously described[26]. Quantification of the immunoprecipitated DNA was performed by qPCR using Thermal Cycler Dice Real Time System TP800 (TAKARA) and SYBR Premix Ex TaqII (TAKARA). The primers for qPCR were as follows: 5′-ATCCTGGGAGTTGGTGATGTC-3′ and 5′-TGCC TCTGAACAGCTCACTT-3′ (for RARβ2 RARE); 5′-AAGTGGAATGTCCCAT CAGCA-3′ and 5′-AGGGCTTCACACCCAGAGCTA-3′ (for RARβ2 distal); 5′-CCACACAGGAGTGGG TGACA-3′ and 5′-ATGGAGGCCTCATGCTCA CA-3′ (for C/EBPε RARE); 5′-TACTGGCAACACAAGCCCAAG-3′ and 5′-AC TGCATTCCAGCCCAGGAG-3′ (C/EBPε distal).

Biochemical purification using the GST-fused RARα LBD. Recombinant GST-fused RARα LBD was expressed in *E. coli*. Nuclear extracts derived from undifferentiated and differentiated HL60 cells were prepared as previously described[49,50]. Nuclear extracts (0.5 g) were incubated overnight (12 h) with a small amount (<2 μl) of glutathione sepharose 4B (GE Healthcare), which was saturated with 5 μg recombinant bait proteins, in the GST purification buffer (50 mM Tris-HCl, 0.1 mM EDTA, 5 mM $MgCl_2$, 0.5 M KCl, 0.1% (v/v) NP-40, 20% (v/v) glycerol, 2 mM DTT, 1 mM benzamidine, 0.2 mM PMSF, pH 7.9), supplemented with or without 1 μM RA. The beads were centrifuged along with 40 μl of additional empty beads and washed with an excess of GST purification buffer. Bound proteins were sequentially eluted three times with 40 μl of GST purification buffer plus 20 mM glutathione. For identification of the polypeptide, 10 μl of the 120 μl eluate was resolved in 4–16% gradient SDS–PAGE. After silver staining with SilverQuest silver staining kit (Invitrogen), visible polypeptides were excised and subjected to tryptic digestion and peptide mass finger print using MALDI–TOF/MS.

Biochemical purification of the Flag–MLL5 complex. HL60 cells stably expressing Flag–MLL5 were established as described in Supplementary Fig. 7 and then cultured in spinner flasks on a 50 l scale. For anti-Flag affinity purification, the prepared nuclear extracts (>0.5 g) were incubated overnight with 50 μl of anti-Flag M2 resin (Sigma) in buffer D (20 mM Tris-HCl, 0.2 mM EDTA, 5 mM $MgCl_2$, 0.1 M KCl, 0.05% (v/v) NP-40, 10% (v/v) glycerol, 0.5 mM DTT, 1 mM benzamidine, 0.2 mM PMSF, pH 7.9). The beads were washed extensively with buffer D and sequentially eluted three times with 50 μl of buffer D plus 0.3 μg ml^{-1} Flag peptide (Sigma). SDS–PAGE and silver-staining analysis was then performed on 2 μl of the 150 μl eluates. For gel filtration, 100 μl of the fractions from the anti-Flag purification was loaded onto a Superose 6 10/300 GL column connected to the AKTA explorer (GE Healthcare) and fractionated with buffer D into 500 μl fractions. For WGA lectin purification, HKMT active fractions (fractions 18–24) from the gel filtration chromatography were further purified and concentrated with a glycoprotein isolation kit, WGA (Pierce), according to the manufacturer's instructions. The protein identification was performed as described previously.

Recombinant proteins. Human *MLL5* cDNA fused with Flag, human *OGT* cDNA fused with 6 × His and mouse *HCF1-N* fused with Flag were subcloned into the pFastBac vector (Invitrogen). Recombinant Flag–MLL5, 6 × His-OGT and Flag–HCF-1N were individually expressed in Sf9 cells as described in a previous report[27]. For purification of Flag–MLL5 and Flag–HCF-1N, whole-cell lysates prepared from the infected cells were combined with TNE350 (TNE supplemented with 350 mM KCl), incubated with anti-Flag M2 resin, washed with TNE, and eluted with TNE supplemented with 0.4 μg ml^{-1} of Flag peptide. For 6 × His–OGT preparation, 6 × His–OGT was partially purified with HIS-Select Nickel Affinity Gel (Sigma) as previously described[27]. The eluates were diluted 1:20 with BC0 (20 mM HEPES, 0.2 mM EDTA, 10% (v/v) glycerol, pH 7.9) and loaded onto a ResourceQ column (BD Healthcare). 6 × His–OGT, fractionated on a linear gradient (0–0.5 M KCl), had a peak at 0.18 M KCl.

***In vitro* HKMT assay (autoradiographic detection).** For substrate preparation, nucleosomes were reconstituted *in vitro* with purified HeLa histone octamers and plasmid by a salt dilution method. In brief, 5 μg of HeLa histone octamers were incubated with 5.5 μg of plasmid DNA and 2 M NaCl for 15 min at 37 °C. The NaCl concentration of the reaction was serially diluted to 1.5, 1, 0.8, 0.7, 0.6, 0.5, 0.4, 0.25 and 0.2 M by adding dilution buffer (50 mM HEPES, 1 mM EDTA, 0.1% NP-40, 5 mM DTT and 0.5 mM PMSF, pH 7.5), with 15 min incubations at 30 °C for each dilution step. The reaction is brought to 0.1 M NaCl by adding a stock buffer (10 mM Tris-HCl, 1 mM EDTA, 0.1% NP-40, 5 mM DTT, 0.5 mM PMSF, 20% glycerol). MLL5-L complex was incubated with 0.5 μg of the various substrates (that is, purified HeLa histone H1, purified HeLa histone octamer and the reconstituted nucleosome), in 20 μl reactions of HKMT buffer (20 mM Tris-HCl, 4 mM EDTA, 1 mM PMSF, 0.5 mM DTT, pH 7.9) with 0.3 μM (1 μCi) *S*-adenosyl-L-(methyl-^{3}H)methionine (GE Healthcare) for 1 h at 30 °C. The HKMT reaction was stopped by directly adding the SDS-sampling buffer. The reactions were resolved with SDS–PAGE, treated with EN^{3}HANCE (NEN Lifescience), and then subjected to autoradiography with Transcreen-LE intensifying screen (Kodak) and BioMax XAR film (Kodak).

***In vitro* HKMT assay (mass spectrometric detection).** The following peptides were used: H3 tail (1–21), ARTKQTARKSTGGKAPRKQLA-GG-biocytin (molecular weight of 2,722.2 Da); K4A, ARTAQTARKSTGGKAPRKQLA (2,197.5 Da). The purified complexes containing MLL5 and its mutants were incubated with 1 μg of the various histone tail peptides in 50-μl reactions (20 mM Tris-HCl, 4 mM EDTA, 1 mM PMSF, 0.5 mM DTT, pH 7.9) with 0.5 mM *S*-adenosyl-L-methionine (Sigma) for approximately 2 days at 30 °C. An aliquot (2 μl) of the HKMT reaction mixture was diluted 50 times with

0.1% formic acid, and loaded on Poros Tip R2 (PE Biosystems) reversed-phase beads packed into an Eppendorf gel-loading tip. The peptides were deionized by 3 washes with 0.1% acetonitrile and eluted with 30 μl of 30% acetonitrile/0.1% formic acid. A fraction (3 μl) of this peptide pool was directly spotted on Prespotted AncorChip (Bruker), dried and subjected to MALDI–TOF/MS analysis using an UltraFlex TOF/TOF instrument (Bruker).

***In vitro* HKMT assay (scintillation detection).** For peptide substrate analysis, the purified MLL5-L complex and the various immunoprecipitants were incubated with 1 μg of the various histone tail peptides in 50 μl reactions of HKMT buffer for 1 h at 30 °C. In the case of the HKMT assay on the immunoprecipitated beads, the reaction was continuously warmed and agitated using BIO Shaker MBR-022 (TAITEC). The reaction was spotted on a P81 paper (Whatman), dried and washed with carbonate buffer (pH 9.2). The spotted P81 paper was then soaked with Atomlight fluid (Packard Biosciences) and subjected to liquid scintillation (Beckman).

***In vitro O*-GlcNAcylation assay.** 0.5 μg of 6 × His–OGT proteins was incubated with 0.5 μg of Flag–MLL5 and 0.2 mM (0.2 μCi) UDP–*N*-acetyl-D-(U-^{14}C)glucosamine (GE Healthcare) in 25 μl reactions (50 mM Tris-HCl, 12.5 mM $MgCl_2$, 1 mM DTT, pH 7.5) for 1 h at 37 °C. The reactions were resolved with SDS–PAGE, treated with EN^{3}HANCE (NEN Lifescience), and then subjected to autoradiography with Transcreen-LE intensifying screen (Kodak) and BioMax XAR film (Kodak).

LETTERS

CtIP–BRCA1 modulates the choice of DNA double-strand-break repair pathway throughout the cell cycle

Maximina H. Yun[1] & Kevin Hiom[1]

The repair of DNA double-strand breaks (DSBs) is tightly regulated during the cell cycle. In G1 phase, the absence of a sister chromatid means that repair of DSBs occurs through non-homologous end-joining or microhomology-mediated end-joining (MMEJ)[1]. These pathways often involve loss of DNA sequences at the break site and are therefore error-prone. In late S and G2 phases, even though DNA end-joining pathways remain functional[2], there is an increase in repair of DSBs by homologous recombination, which is mostly error-free[3,4]. Consequently, the relative contribution of these different pathways to DSB repair in the cell cycle has a large influence on the maintenance of genetic integrity. It has remained unknown how DSBs are directed for repair by different, potentially competing, repair pathways. Here we identify a role for CtIP (also known as RBBP8) in this process in the avian B-cell line DT40. We establish that CtIP is required not only for repair of DSBs by homologous recombination in S/G2 phase but also for MMEJ in G1. The function of CtIP in homologous recombination, but not MMEJ, is dependent on the phosphorylation of serine residue 327 and recruitment of BRCA1. Cells expressing CtIP protein that cannot be phosphorylated at serine 327 are specifically defective in homologous recombination and have a decreased level of single-stranded DNA after DNA damage, whereas MMEJ remains unaffected. Our data support a model in which phosphorylation of serine 327 of CtIP as cells enter S phase and the recruitment of BRCA1 functions as a molecular switch to shift the balance of DSB repair from error-prone DNA end-joining to error-free homologous recombination.

Because non-homologous end-joining (NHEJ), MMEJ and homologous recombination pathways all operate in S and G2 phases, the repair of a DSB by a particular pathway is not determined simply by limiting the availability of specific repair factors to defined phases of the cell cycle. We reasoned that this choice is more likely to be determined by proteins that function in the initial steps of DSB repair, involving the recognition and processing of DNA ends (Supplementary Fig.1). Moreover, we predicted that a potential candidate for this role is CtIP, which is known to promote resection of DNA ends to the single-stranded DNA (ssDNA) tails that are essential for homologous recombination[5,6].

Analysis of CtIP in mammalian cells is difficult because homozygous deletion of *CtIP* in mice results in early embryonic lethality[7]. To circumvent this problem, we generated *CtIP* null mutant cells from the avian B-cell line DT40. In DT40 *CtIP* is present in three copies owing to a chromosomal duplication of chromosome 2 (ref. 8). We disrupted all three *CtIP* alleles, deleting exons 1 and 2, which contained the initiation codon and 5′ untranscribed region of the gene, and replaced them with antibiotic-resistance cassettes (Supplementary Fig. 2a). We confirmed the generation of $CtIP^{-/-/-}$ mutant cells by Southern blot analysis and the loss of CtIP protein expression by western blot (Supplementary Fig. 2b, c).

In common with other DNA-repair-defective mutants[9,10], $CtIP^{-/-/-}$ cells have a reduced proliferation rate compared with wild-type cells (Supplementary Fig. 3a). Moreover, in clonogenic survival assays they are highly sensitive to X-rays, which cause DSBs (Fig. 1a). They are also sensitive to cisplatin, which generates interstrand DNA crosslinks that may also lead to the generation of DSBs during replication (Fig. 1a). In contrast, $CtIP^{-/-/-}$ cells are not very sensitive to ultraviolet light, which causes pyrimidine dimers and 6-4 photoproducts[11] (Fig. 1a). Importantly, expression of human CtIP in $CtIP^{-/-/-}$ mutant cells fully restored the resistance of these cells to X-rays, confirming that the human and avian CtIP proteins are functionally conserved (Fig. 1a and Supplementary Fig. 2d).

It was reported that CtIP promotes resistance to DSB-inducing agents exclusively during S and G2 phases[5,12]. Accordingly, $CtIP^{-/-/-}$ cells isolated by elutriation in S/G2 phase (Supplementary Fig. 4) are fivefold to sixfold more sensitive to X-ray damage (lethal dose at which there is 10% survival, $LD_{10} = 2.5$ Gy) than wild-type cells, and approximately twofold more sensitive than NHEJ-defective *KU70* (also known as *XRCC6*) mutant cells (Fig. 2a, upper panel). Surprisingly, $CtIP^{-/-/-}$ cells isolated in G1 phase are also sensitive to X-ray-induced DNA damage ($LD_{10} = 2.2$ Gy) (Fig. 2a, lower panel), indicating that CtIP function is not limited to S/G2 phase, as previously proposed, but contributes to the repair of DSBs throughout the cell cycle. Also, because homologous recombination does not function in G1 phase, CtIP must be involved in a second pathway for repairing DSBs.

The fact that CtIP is required for repair of X-ray-induced DSBs in G1 and S/G2 phases raised the possibility that it functions commonly in both homologous recombination and DNA end-joining. To address this we made use of several green fluorescent protein (GFP) reporter assays that measure the repair of a restriction-enzyme-induced DSB by different repair pathways (Supplementary Fig. 5).

For homologous recombination we measured the repair of a defined I-SceI-induced genomic DSB in a defective GFP reporter gene (Sce−GFP[13]; Supplementary Fig. 5). In line with previous studies[7], we observed a tenfold defect in homologous recombination for $CtIP^{-/-/-}$ mutant cells compared to wild-type cells (Fig. 3a). A second form of homology-directed DNA repair is single-strand annealing (SSA). This occurs when a DSB is generated between two directly repeated sequences and is achieved by resection of DNA ends to produce homologous ssDNA tails that can anneal to promote joining[14]. Again we found that $CtIP^{-/-/-}$ mutant cells are tenfold defective in SSA compared with wild-type cells (Fig. 3b). We conclude that CtIP has an important role in DSB repair by homologous recombination and that this accounts for the sensitivity of $CtIP^{-/-/-}$ cells to X-ray damage in S and G2 phases.

DNA end-joining is achieved either through NHEJ, whereby broken DNA ends are directly rejoined, or by MMEJ, in which DNA ends are resected locally to reveal short regions of complementary DNA (4 to 6 nucleotides) that stabilize broken ends for ligation[1]. We tested

[1]Division of Protein and Nucleic Acid Chemistry, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK.

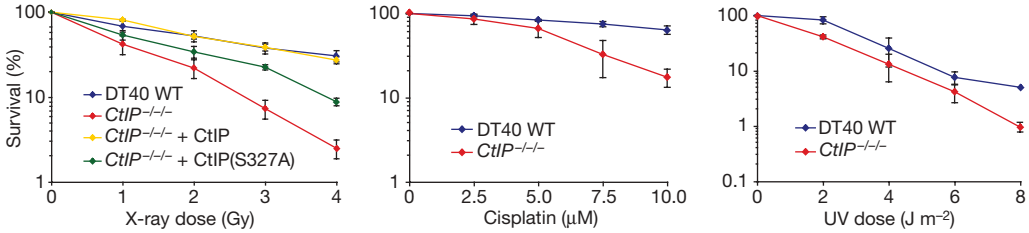


Figure 1 | Sensitivity of *CtIP* $^{-/-/-}$ mutant cells to DNA-damaging agents. Clonogenic survival assays with asynchronous cell populations after exposure to X-rays (left), cisplatin (middle) and ultraviolet light (UV, right). Data are the mean of three independent experiments; error bars, s.d.

CtIP $^{-/-/-}$ cells for both types of end-joining and found no defect in accurate NHEJ (Fig. 3c). In fact, we observed a slight increase in NHEJ activity in the *CtIP* $^{-/-/-}$ mutant compared with wild-type cells. In contrast, for MMEJ we observed a fourfold to fivefold defect in the *CtIP* $^{-/-/-}$ mutant compared with wild-type cells (Fig. 3d).

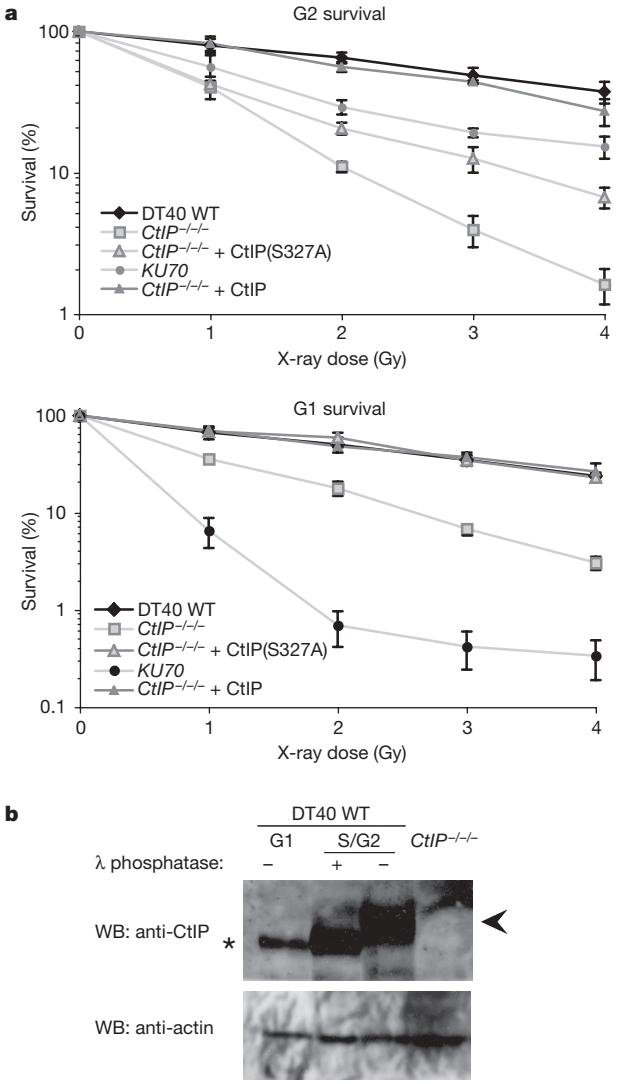


Figure 2 | *CtIP* $^{-/-/-}$ mutant cells are sensitive to X-rays in both G1 and S/G2 phases of the cell cycle. **a**, Clonogenic colony survival assay after exposure of cell lines to X-rays, synchronized by elutriation in either S/G2 (top) or G1 (bottom) stages of the cell cycle. Data represent three independent experiments; error bars, s.d. **b**, Western blot (WB) showing the presence of CtIP in G1 and S/G2. Phosphorylated CtIP is indicated with an arrowhead. Only unphosphorylated CtIP (asterisk) is seen in G1. Whole-cell extracts were prepared from elutriated cell populations. Where indicated, S/G2 cell extract were treated with λ phosphatase for 2 h at 30 °C. Western blot with an antibody against actin was used as the protein-loading control.

In a second assay we transfected cells with linearized plasmid, recovered the repaired plasmids after 24 h and examined the DNA sequences surrounding the joints (Supplementary Fig. 6). Both wild-type and *CtIP* $^{-/-/-}$ mutant cells repaired most breaks through a combination of accurate and inaccurate NHEJ (Fig. 3e). Where NHEJ was inaccurate, the spectrum of deletions at the break site was similar in *CtIP* $^{-/-/-}$ and wild-type cells. Nevertheless, in the *CtIP* $^{-/-/-}$ mutant cells we again detected a reduction in MMEJ (Fig. 3e).

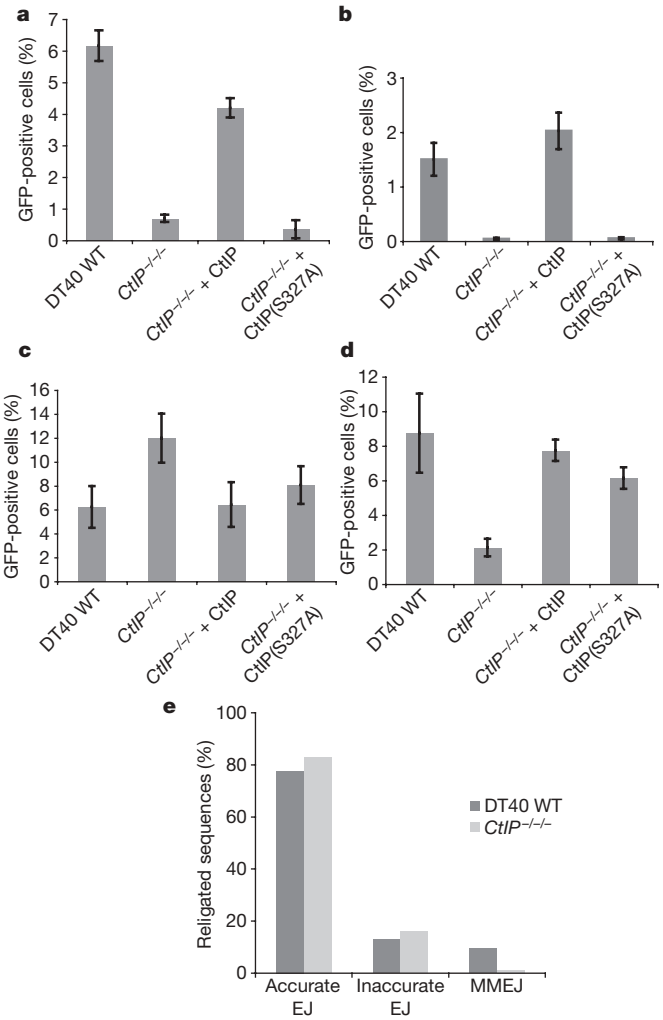


Figure 3 | *CtIP* $^{-/-/-}$ mutant cells are defective for homologous recombination and MMEJ. Repair is indicated by percentage of cells expressing GFP as described in Supplementary Fig. 5. **a**–**e**, Repair by homologous recombination (**a**); single-strand annealing (**b**); accurate NHEJ (**c**); MMEJ (**d**); analysis of DNA sequences at repaired break sites in wild-type and *CtIP* $^{-/-/-}$ mutant cells (**e**). Individual sequences (shown in Supplementary Fig. 4) were classified according to the nature of their joints into 'accurate end-joining (EJ)', 'inaccurate EJ' and 'MMEJ'. Data are the mean of three independent experiments; error bars, s.d.

Together with observations in 293 cells that short-interfering-RNA-mediated knockdown of CtIP alters the balance of DSB repair[15], these data establish a role for CtIP in MMEJ and provide an explanation for the defects in DSB repair observed in *CtIP* $^{-/-/-}$ cells during G1 phase.

A role for CtIP in DSB repair during G1 was unexpected. Previous studies indicated that CtIP is present at very low levels outside of the S and G2 phases[12,16]. Nevertheless, CtIP was present in extracts from DT40 cells in G1 albeit at reduced levels compared to cells in S/G2 (Fig. 2b). Furthermore, whereas in G1 phase CtIP is largely unmodified, most CtIP in S/G2 is phosphorylated (Fig. 2b).

Previously it was demonstrated that CtIP is phosphorylated on serine residue 327 as cells enter S phase, which mediates its interaction with the tumour suppressor BRCA1 that is required for the transient G2/M checkpoint[12,17]. Therefore, we next considered whether phosphorylation of serine 327 might also regulate the function of CtIP in DSB repair. To investigate this we expressed a mutant form of human CtIP, in which serine 327 was substituted by alanine (human CtIP(S327A)), in *CtIP* $^{-/-/-}$ cells and examined its sensitivity to X-rays. Although expression of human CtIP(S327A) improved the survival of *CtIP* $^{-/-/-}$ cells to X-rays, complementation was only partial (Fig. 1a), indicating that mutation of serine 327 results in loss of some but not all of the repair functions of CtIP.

The picture became clearer when we looked at the survival in response to X-ray damage in different phases of the cell cycle. We found that expression of either human CtIP or human CtIP(S327A) in *CtIP* $^{-/-/-}$ cells fully restored resistance to X-rays in G1 phase and restored MMEJ to wild-type levels in a plasmid assay (Fig. 3d), indicating that phosphorylation on serine 327 is not required for the repair of DSBs by MMEJ (Fig. 2a). However, expression of human CtIP(S327A) in *CtIP* $^{-/-/-}$ cells did not restore homologous recombination, indicating that phosphorylation of serine 327 is important for this function (Fig. 3a, b). Accordingly, the sensitivity of *CtIP* $^{-/-/-}$ cells to X-rays in S/G2 phase was only partially restored by human CtIP(S327A) (Fig. 2a), which can be accounted for by restoration of MMEJ, but not homologous recombination, in these cells.

We investigated how phosphorylation of serine 327 on CtIP might increase the contribution of homologous recombination during S phase. It is known that CtIP promotes the resection of DNA ends[5] and that this is an important step in both homologous recombination and MMEJ. However, because MMEJ requires small local regions of microhomology it is likely that the resection required for this pathway is less extensive than for strand exchange in homologous recombination. We reasoned, therefore, that phosphorylation of CtIP might upregulate the generation of ssDNA.

To test this we took advantage of the fact that 5-bromodeoxyuridine (BrdU) incorporated into the genome is detectable by anti-BrdU antibody only in regions of ssDNA[18]. We cultured cells in BrdU, treated them with X-rays and stained for BrdU at intervals up to 2 h (Fig. 4a). Approximately 25% of unirradiated cells exhibit ten or more BrdU foci. After exposure to X-rays (8 Gy) we observed a time-dependent increase in BrdU staining in wild-type cells until, after 2 h, approximately 50% contained ten or more BrdU foci. Over the same time period only 30% of *CtIP* $^{-/-/-}$ mutant cells stained with BrdU, indicating that, although these mutant cells are not completely defective in the generation of ssDNA after exposure to X-rays, it occurs more slowly. Moreover, expression of human CtIP in the *CtIP* $^{-/-/-}$ fully rescued this delay. In contrast, expression of human CtIP(S327A) cells did not complement this defect in the *CtIP* $^{-/-/-}$ mutant, indicating that the DNA-damage-dependent increase in ssDNA is linked to the phosphorylation of serine 327. The delayed generation of ssDNA is not linked to reduced growth rate because cells expressing human CtIP(S327A) exhibit reduced BrdU foci but proliferate normally (Supplementary Fig. 7).

Our data place CtIP at the 'crossroads' between DNA end-joining and homologous recombination pathways for the repair of DSBs, with phosphorylation of serine 327 acting as a cell-cycle-dependent switch that regulates CtIP-dependent DNA end resection.

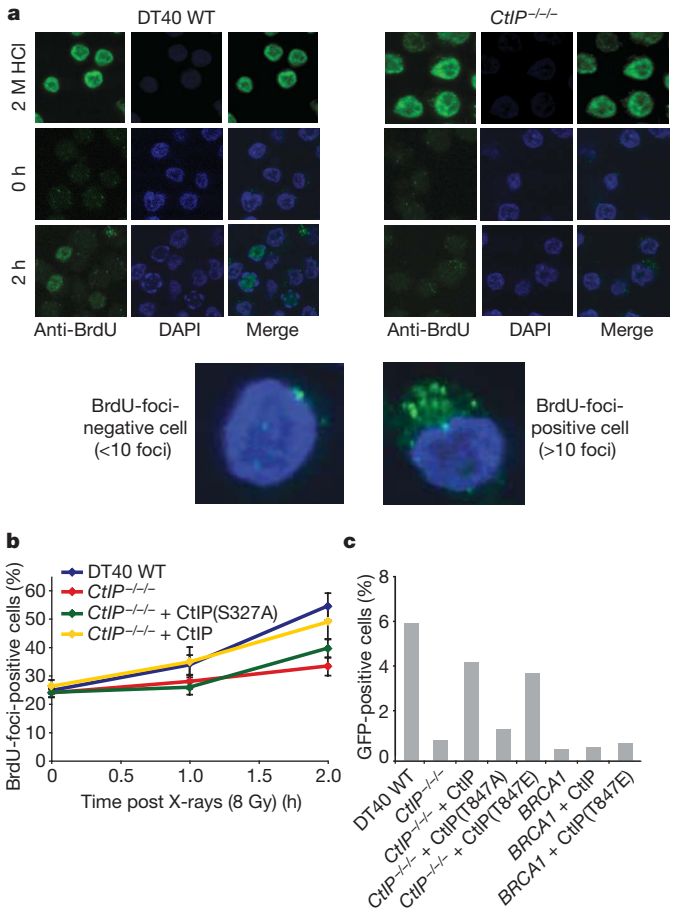


Figure 4 | Phosphorylation of serine 327 is required for generation of ssDNA in DT40 cells. a, Cells were labelled with 20 μM BrdU and treated with 8 Gy X-rays (where indicated). ssDNA was detected over time by staining with antibody against BrdU and analysed using confocal microscopy. Staining was performed under DNA-denaturing (2 M HCl) or native conditions as indicated. Original magnification, ×1,000. **b**, Kinetics of ssDNA generation from **a**. Cells containing ten or more foci were scored as positive. **c**, Homologous recombination assay, performed as in Fig. 3 with indicated cell lines. This shows that expression of 'constitutively activated' CtIP(T847E) restores homologous recombination in *CtIP* $^{-/-/-}$ mutant cells but not in *BRCA1* cells. Data are the mean of three independent experiments; error bars, s.d. 4,6-diamidino-2-phenylindole (DAPI).

Phosphorylation of serine 327 is known to control the interaction of CtIP with BRCA1 in DT40 (ref. 12; Supplementary Fig. 8a). Moreover, like serine 327 of CtIP, BRCA1 is required for repair of DSBs by homologous recombination but not MMEJ (Fig. 4c and Supplementary Fig. 8b), indicating that the recruitment of BRCA1 to CtIP may be a determining factor in this switch.

Serine 327 lies within a weak cyclin-dependent kinase (CDK) consensus site (SP/TP) in CtIP, which, combined with the presence of a cyclin interaction motif (ZRXL)[19], makes it a likely substrate for CDKs. This residue is specific to CtIP in complex eukaryotes and is not present in Sae2, the CtIP-like protein of the budding yeast *S. cerevisiae*. In yeast, homologous recombination and DNA end resection are promoted by CDK-dependent phosphorylation of Sae2 on serine 267 (refs 20 and 21). Interestingly, this site is conserved at threonine residue 847 of vertebrate CtIP[21], which we show here performs a similar function to serine 267 of Sae2 (Supplementary Figs 9 and 10). In CtIP, substitution of threonine 847 with alanine causes defects in the repair of DSBs by homologous recombination in S phase, but does not affect MMEJ. Accordingly, cells expressing CtIP(T847A) have increased mutagenic repair of DSBs (Supplementary Fig. 10d). Moreover, as in yeast, the requirement for phosphorylation of threonine 847 can be circumvented by the use of a phosphomimic mutation

in which threonine 847 in CtIP is replaced by glutamate (CTIP(T847E)). This mutant exhibits normal resection and restoration of homologous recombination without phosphorylation at threonine 847 (Supplementary Figs 10 and 11).

Although in yeast the switch to accurate DSB repair in S phase is controlled by phosphorylation of a single CDK site in Sae2, our data demonstrate that phosphorylation of CtIP at two independent CDK sites (serine 327 and threonine 847) is required in vertebrates. The particular importance of serine 327 and the requirement of BRCA1 for this switch were established in two ways. First, although expression of CtIP(T847E) restores homologous recombination to *CtIP* $^{-/-/-}$ mutant cells without the requirement for phosphorylation at residue 847, we found that the CtIP(S327A,T847E) double mutant, in which phosphorylation at serine 327 is not possible, is defective in homologous recombination (Supplementary Fig. 10e). Second, we show that a CtIP(T847E) mutant does not restore homologous recombination to *BRCA1* mutant cells, confirming that recruitment of BRCA1 by CtIP is required for efficient homologous recombination function independently from the activation at threonine 847 (Fig. 4c).

Together these data establish a pivotal role for CtIP and BRCA1 in a switch that has profound consequences for the maintenance of genetic integrity in DNA-damaged cells by facilitating a shift from predominantly error-prone repair of DSBs by DNA end-joining in G1 to the accurate repair afforded by homologous recombination in S and G2 phases.

METHODS SUMMARY

DT40 chicken cells were propagated in standard RPMI supplemented media[11] at 37 °C, 6% CO_2. Transfections were carried out by electroporation as previously described[11]. For the generation of CtIP knockout cells, genomic DNA sequences were amplified with LA-TaKara and targeting vectors assembled (neomycin, histidinol and blasticidin). For each transfection, 30 μg of the selected vector was linearized using NotI (NEB) and transfected into DT40 cells. Clones were obtained after 2−4 weeks of growth under selective media. Genomic DNA from individual clones was prepared with PURAGENE DNA Purification Kit (Gentra). Twenty microlitres of each was digested with PacI and XmnI (NEB) at 37 °C overnight (12–16 h), and used in Southern blotting to detect targeted integration. Membranes were hybridized with a ^{32}P GgCtIP probe (Supplementary Fig. 2b). Primers used for generating targeting vectors are listed in Methods. Expression vectors for complementation studies were generated by insertion of *CtIP* complementary DNA into pCR2.1-TOPO and subsequent subcloning into pcDNA3.1/zeo(+). Primers used for cloning CtIP are listed in Methods. Clonogenic survival assays were performed as previously described[11]. Cells were isolated at specific stages of the cell cycle by elutriation. In brief, 10^9 DT40 cells were collected by centrifugation, resuspended in 20 ml RPMI medium and drawn into a standard (4 ml) chamber at 3,000*g* in a Beckman JE-5.0 centrifugal elutriation rotor. Cells were maintained at a constant flow rate (38 ml min^{-1}) and were elutriated by decreasing the rotor speed at intervals of 11*g*. For each elutriation interval, a 150 ml fraction was collected and cells spun down, resuspended in 3 ml cold PBS and counted. An aliquot of 1×10^5 cells was incubated for 10 min with 1 μl of the cell-permeable DNA dye Draq5 (Biostatus Ltd) and the cell cycle profile determined by FACS analysis on a FACScan cytometer (Becton Dickinson).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 28 May 2008; accepted 3 March 2009.
Published online 8 April 2009.

1. Ma, J. L., Kim, E. M., Haber, J. E. & Lee, S. E. Yeast Mre11 and Rad1 proteins define a Ku-independent mechanism to repair double-strand breaks lacking overlapping end sequences. *Mol. Cell. Biol.* **23,** 8820–8828 (2003).
2. Kim, J. S. *et al.* Independent and sequential recruitment of NHEJ and HR factors to DNA damage sites in mammalian cells. *J. Cell Biol.* **170,** 341–347 (2005).
3. Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J.* **17,** 5497–5508 (1998).
4. Pâques, F. & Haber, J. E. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63,** 349–404 (1999).
5. Sartori, A. A. *et al.* Human CtIP promotes DNA end resection. *Nature* **450,** 509–514 (2007).
6. Baumann, P. & West, S. C. Role of the human RAD51 protein in homologous recombination and double-stranded-break repair. *Trends Biochem. Sci.* **23,** 247–251 (1998).
7. Chen, P. L. *et al.* Inactivation of CtIP leads to early embryonic lethality mediated by G1 restraint and to tumorigenesis by haploid insufficiency. *Mol. Cell. Biol.* **25,** 3535–3542 (2005).
8. Sonoda, E., Morrison, C., Yamashita, Y. M., Takata, M. & Takeda, S. Reverse genetic studies of homologous DNA recombination using the chicken B-lymphocyte line, DT40. *Phil. Trans. R. Soc. Lond. B* **356,** 111–117 (2001).
9. Tauchi, H., Matsuura, S., Kobayashi, J., Sakamoto, S. & Komatsu, K. Nijmegen breakage syndrome gene, *NBS1*, and molecular links to factors for genome stability. *Oncogene* **21,** 8967–8980 (2002).
10. Simpson, L. J. & Sale, J. E. Rev1 is essential for DNA damage tolerance and non-templated immunoglobulin gene mutation in a vertebrate cell line. *EMBO J.* **22,** 1654–1664 (2003).
11. Bridge, W. L., Vandenberg, C. J., Franklin, R. J. & Hiom, K. The BRIP1 helicase functions independently of BRCA1 in the Fanconi anemia pathway for DNA crosslink repair. *Nature Genet.* **37,** 953–957 (2005).
12. Yu, X. & Chen, J. DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. *Mol. Cell. Biol.* **24,** 9478–9486 (2004).
13. Pierce, A. J., Johnson, R. D., Thompson, L. H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13,** 2633–2638 (1999).
14. Stark, J. M., Pierce, A. J., Oh, J., Pastink, A. & Jasin, M. Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Mol. Cell. Biol.* **24,** 9305–9316 (2004).
15. Bennardo, N., Cheng, A., Huang, N. & Stark, J. M. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS Genet.* **4,** e1000110 (2008).
16. Yu, X., Wu, L. C., Bowcock, A. M., Aronheim, A. & Baer, R. The C-terminal (BRCT) domains of BRCA1 interact *in vivo* with CtIP, a protein implicated in the CtBP pathway of transcriptional repression. *J. Biol. Chem.* **273,** 25388–25392 (1998).
17. Greenberg, R. A. *et al.* Multifactorial contributions to an acute DNA damage response by BRCA1/BARD1-containing complexes. *Genes Dev.* **20,** 34–46 (2006).
18. Schlegel, B. P., Jodelka, F. M. & Nunez, R. BRCA1 promotes induction of ssDNA by ionizing radiation. *Cancer Res.* 66, 5181–5189 (2006.
19. Endicott, J. A., Noble, M. E. & Tucker, J. A. Cyclin-dependent kinases: inhibition and substrate recognition. *Curr. Opin. Struct. Biol.* **9,** 738–744 (1999).
20. Ira, G. *et al.* DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. *Nature* **431,** 1011–1017 (2004).
21. Huertas, P., Cortes-Ledesma, F., Sartori, A. A., Aguilera, A. & Jackson, S. P. CDK targets Sae2 to control DNA-end resection and homologous recombination. *Nature* **455,** 689–692 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements The authors would like to thank M. Jasin for gifts of DR–GFP and pHPRT-SSA-GFP, S. Takeda for the gift of *KU70* DT40, and J. Di Noia for DT40 DTDR-17. We would also like to thank our colleagues C. Rada and J. Sale for comments and suggestions during the preparation of this manuscript. M.H.Y. is a Milstein Student of the Darwin Trust, Edinburgh, Scotland.

Author Contributions All experiments were performed by M.H.Y. and were conceived by M.H.Y. and K.H. K.H. and M.H.Y. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.H. (hiom@mrc-lmb.cam.ac.uk).

METHODS

GgCtIP knockout primers. For the construction of the targeting vectors the following primers were used. Left arm: 5′SalI-CtIP L arm, 5′-gtcgacCAGAATCC CTCAGGTCCTGGAATG-3′; 3′BglII-CtIP L arm, 5′-AGATCTATTTGTGGTC TGCTGGCTGTTGGG-3′. Right arm: 5′BglII-CtIP R arm, 5′-agatctTGACCAT TTTGGTCCAGTTAAATC-3′; 3′NotI-CtIP R arm, 5′-GCGGCCGCGCACTC TATTGGAGGTATTGCC-3′. Restriction sites are underlined. For the screening of clones by Southern blot, a probe was designed with the following primers. Probe forward: 5′-CCTGATGAATGTCTACTGTGGTGCATTCTT-3′. Probe reverse: 5′-CGCCTTTAATAAGTTAGATCATCAGTATGA-3′.

pCMV/I-Sce1/GFP primers. EJfwd: 5′-CTAGGGATAACAGGGTAATCGGC TAG-3′. EJrev: 5′-CCATTACCCTGTTATCCCTAGCTAG-3′.

CtIP cloning primers. 5′BamHI CtIP (pcDNA3.1): 5′-GGATCCACCATGAAC ATCTTGGGAAGCAGCTGTG-3′. 3′NotI CtIP (pcDNA3.1): 5′-GCGGCCGCC TATGTCTTCTGCTCCTTGCCT-3′.

Primers for site-directed mutagenesis of CtIP. CtIP(S327A): forward primer, 5′-CCTACTCGAGTGTCAGCTCCTGTATTTGGAG-3′; reverse primer, 5′-CT CCAAATACAGGAGCTGACACTCGAGTAGG-3′. CtIP(T847A): forward primer, CTACATTCCACCCAACGCACCAGAGAATT; reverse primer, AATTCT CTGGTGCGTTGGGTGGAATGTAG. CtIP(T847E): forward primer, CTACA TTCCACCCAACGAACCAGAGAATT; reverse primer, AATTCTCTGGTTCG TTGGGTGGAATGTAG.

Inmunoprecipitation and western blotting. Whole-cell extracts were prepared from 15×10^6 chicken DT40 cells lysed in NET-N buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.1% NP-40) supplemented with protease and phosphatase inhibitors (1 mM NaF, 1 mM Na_3VO_4 and 10 mM β-glycerophosphate). Nuclear extracts were prepared as previously described.

For immunoprecipitation analysis, nuclear or whole-cell extracts were incubated with 30−10 μl of the indicated antibodies and 40 μl of protein A/G-Sepharose beads (Amersham). Beads were collected by centrifugation, washed 5 times with 1 ml NET-N buffer and boiled for 5 min in 50 μl of 2× SDS loading buffer (0.1 mM Tris-HCl, pH 6.8, 4% SDS, 20% glycerol, 0.1% β-mercaptoethanol, 0.004% bromophenol blue). Twenty microlitres of each sample were resolved by SDS–PAGE. Proteins were identified by a standard western blot using the following primary antibodies: CtIP (T-16) sc-5970 (Santa Cruz Inc.) 1:500 dilution, GgBRCA1 (non commercial) 1:1,000 dilution. Proteins were visualized using ECL-plus (Amersham).

Cell cycle analysis. Cells were pulsed with 20 μM BrdU (Sigma) for 20 min, washed twice with PBS and fixed overnight in 70% ethanol at −20 °C. Cells were stained against BrdU and propidium iodide, and cell cycle analysis performed as described[10].

Homologous recombination and SSA assays. For the homologous recombination assay, DT40 cells stably expressing a single copy of DR−GFP (M. Jasin, Sloan-Kettering) were a gift from J. Di Noia. For the SSA assay, the vector pHPRT-SSA-GFP (M. Jasin, Sloan-Kettering) was stably transfected into DT40 or *CtIP* $^{-/-/-}$ mutant cells. The clones obtained after puromicin selection were screened for single integrants by Southern blotting. To perform each of the experiments, cells were transfected with either pCASce[13] or the control plasmid pDsRed1-N1 (Clontech) using the Amaxa nucleofection system. They were then suspended in 5 ml of medium, incubated at 37 °C for 24 h and washed in 500 μl of PBS. 8 μg ml^{-1} propidium iodide was added to the samples transfected with pCASce. Cells expressing GFP were quantified using a FACScan cytometer (Becton Dickinson). Data are corrected for transfection efficiency as measured by the percentage of cells that express DsRed.

MMEJ and end-joining assays. For the MMEJ assay, the vector pCMV/cyto/myc/GFP (Invitrogen) was modified by insertion of a pre-annealed oligonucleotide (EJfwd and EJrev) carrying an I-SceI restriction site into the unique BmtI site present in the GFP coding sequence to generate pCMV/ I-Sce1/GFP. The vector was digested with I-SceI at 37 °C overnight, analysed for complete digestion and ethanol precipitated. Ten micrograms of either uncut pCMV/cyto/myc/GFP (transfection control) or I-SceI-digested pCMV/I-Sce1/GFP was used for transient transfection of DT40 cells with the Amaxa nucleofection system and analysed as described previously. The same protocol was used for the end-joining assay but transfections were carried out with HincII-digested pCMV/cyto/myc/GFP or its uncut version. For the analysis of joined DSBs, cells were transfected with an I-SceI linearized vector pCMV/I-Sce1/GFP as described. Plasmid DNA was extracted and used to tranform DH5 alpha bacteria. Resistant colonies were grown in 96-well agar plates and sequence analysis performed using a pCMV-F primer (GATC Biotech).

DT40 immunofluorescence. DT40 cells were cultured for one and a half cell cycles in the presence of BrdU (20 μM, Invitrogen). After this period, cells were grown in 6-well plates at 3×10^6 cells per well and exposed to X-rays as described previously. Glass coverslips coated with poly-L-lysine (Sigma) were added to each sample. At different times after irradiation, cells were washed with PBS, fixed in 4% paraformaldehyde for 5 min and washed with PBS a second time. Cells subject to denaturing conditions were treated for 30 min with 2 ml of 2 M HCl/0.5% Triton X-100 and neutralized with 2 ml sodium tetraborate, after which they were blocked in PBSTS/1% BSA alongside cells subject to native conditions. Samples were next incubated with mouse polyclonal antibody against BrdU (BD Pharmigen) for 2 h. Coverslips were washed 2× with PBSTS/1% BSA, incubated with an antibody against mouse immunoglobulin, FITC-conjugated (BD Pharmigen) for 1 h, washed twice in PBS and finally mounted on slides with mounting medium containing 4,6-diamidino-2-phenylindole (DAPI, Vectashield). Samples were analysed using a Nikon confocal microscope.

NEWS

Ontario's attractive prospect

A Can$100-million (US$85-million) research fund aims to attract some of the world's leading researchers to Ontario — and keep them there. Launched by the Ontario Ministry of Research and Innovation, the Global Leadership Round in Genomics and Life Sciences fund will finance proteomics, stem-cell and genomics research, with a focus on human health. It may also support genetics and genomics research related to agriculture, environmental protection and clean technologies.

John Wilkinson wants to retain scientists.

"With the mega-stimulus package down south [in the United States], scientists that are not finding the opportunity to build their careers or support their teams are going to look at opportunities south of the border," says Tom Hudson, president and scientific director of the Ontario Institute for Cancer Research in Toronto. John Wilkinson, the Ontario minister of research and innovation, wants the fund to help retain scientists, he said in a statement. Ontario is Canada's largest biomedical research centre and the fourth largest in North America.

Winning projects are likely to be announced in early 2010. The funds will cover salaries, equipment and other research costs, and collaborations with international partners will receive priority. Winners will receive at least Can$3.5 million and up to one-third of the project's cost, with the balance coming from institutional and private-sector partners. Researchers have until the end of August to apply for the peer-reviewed competition.

Scientists say the new funding is timely and could help offset cuts at the federal level. In January, the federal budget called for Canada's three granting councils to scale back their budgets by Can$148 million over three years, starting this year (see *Nature* **457,** 646; 2009). It also failed to provide support to Genome Canada, a not-for-profit funding organization. In April, Genome Canada withdrew Can$18 million from the International Regulome Consortium, a Can$80 million Canadian-led research programme (see *Nature* **458,** 819; 2009).

The government has invested millions in science infrastructure, but many say that's not enough. "Canada has been overbuilding science infrastructure and under-supporting researchers and other workers," says Paul Hebert, director of the Biodiversity Institute of Ontario at the University of Guelph. Hebert, who is the driving force behind the International Barcode of Life project that is cataloguing genetic signatures, calls it an "important" and "timely" funding competition.

The fund will cover salaries of up to Can$20,000 for master's- and PhD-level graduate students and Can$50,000 for postdoctoral fellows. "The best way to train postdocs and students is by having them be part of cutting-edge projects," says Hudson. ■

Hannah Hoag

POSTDOC JOURNAL

Interdisciplinary images

Live jazz music crashes through the dark bar as my friend and I discuss what it means to be 'interdisciplinary'. My beer-loosened words are barely audible over the music.

First, I say, we must define a discipline. I describe my experience as a postdoc. In my first position, I did experimental work in single-molecule biophysics, manipulating muscle proteins with laser tweezers. Perhaps a dozen labs use this method; they have a vocabulary and notation convention that define the discipline.

My current position, I contend, is interdisciplinary. I use mathematical and mechanical models to understand biology. Some of this work addresses the same questions as my previous experimental work. However, for my current work, no single community of researchers exists. One community understands our mathematical methods but not the biological systems; another is familiar with the biology but not the maths.

Conducting research that spans the expertise of two or more groups, yet is understandable to each, presents a challenge. Because researchers are often experts in a single discipline, reviewing interdisciplinary papers can be difficult. Yet it presents the fantastic opportunity to start a new discipline.

As the music swallows my final words, I look across the table at my friend, a new graduate student, and doubt if I have been much help. I wonder how she will define interdisciplinary when she becomes a postdoc. ■

Sam Walcott is a postdoc in theoretical biophysics at Johns Hopkins University in Baltimore, Maryland.

IN BRIEF

Non-tenure on the rise

The proportion of full-time tenured and tenure-track faculty in the United States has declined since the late 1990s and higher-education institutions are relying more on non-tenured 'contingent' faculty members, according to a report by the American Federation of Teachers. The higher-education instructional workforce grew in the past decade along with rising enrolments, the report finds. But colleges and universities have increasingly turned to part-time and full-time non-tenured faculty. From 1997 to 2007, full-time tenured and tenure-track faculty members declined from about a third of the instructional staff to slightly more than a quarter. A similar change was found in all sectors of higher education.

Winding up

A wind-technology testing centre in Boston, Massachusetts, is poised to become a magnet for engineering and technology research and development posts, according to Robert Keough, spokesman for the Massachusetts executive office of energy and environmental affairs. On 12 May, the US Department of Energy announced a $25-million award in economic stimulus funding to the state to accelerate development of the centre. The Boston centre — the nation's first large-blade test facility — will analyse commercial-sized wind turbine blades more than 50 metres and up to 90 metres in length, Keough says. Ten technicians will work at the centre when construction is complete in 2010 and more job opportunities are likely, he says.

Biotech funding rethink

The biotechnology industry performed well in 2008 despite the worldwide economic downturn, but financial services firm Ernst & Young warns that the industry has to find new ways to do business. Revenues of publicly traded biotech companies rose by an average of 12% worldwide in 2008. But capital raised in Europe and the Americas was down by almost half from 2007, and funds raised through initial public offerings fell by 95% for the period (see *Nature* 458, 1062; 2009). "Firms will need to establish more durable models for funding innovation," says Glen Giovannetti, Ernst & Young's global biotechnology leader, on the company's website.

IMAGES.COM/CORBIS

Unmasking the impostor

Feelings of inadequacy in one's field sometimes plague even the most accomplished scientists, especially women. **Karen Kaplan** analyses this apparent phenomenon and its impact.

It usually happens to Cherry Murray when she is about to write a paper or give a talk on a new finding or discovery. The thoughts come unbidden, hammering inside the physicist's head: 'I can't do this. I haven't done enough experiments. I haven't got enough data. I can't write the paper well enough yet or give the talk.'

These aren't the routine self-doubts of a young researcher. Murray is principal associate director for science and technology at Lawrence Livermore National Laboratory (LLNL) in California, and president of the American Physical Society. On 1 July she will become dean of Harvard University's School of Engineering and Applied Sciences. Referring to her work, she says: "I have to tell myself this doesn't have to be perfect."

What Murray describes — an overwhelming sense of being a fraud, a phony, of not being good enough for her job, despite much evidence to the contrary — was first identified more than 30 years ago by two clinical psychologists who dubbed it the 'impostor phenomenon' (IP) (P. Clance and S. Imes *Psychother. Theor. Res.* **15,** 241–247; 1978). In their paper, Pauline Clance and Suzanne Imes describe women who, despite reaching significant intellectual milestones ranging from advanced degrees to professional awards, cannot internalize their success or convince themselves they deserve it. "They consider themselves to be 'impostors'," wrote Clance and Imes. "Numerous achievements, which one might expect to provide ample objective evidence of superior intellectual functioning, do not appear to affect the impostor belief."

Before taking up her LLNL post, Murray spent decades at Bell Laboratories and left as senior vice-president for physical sciences and wireless research. She has published more than 70 papers in peer-reviewed journals, has won a number of awards, holds two patents and has served on more than 80 national and international scientific committees and governing boards. Yet the self-doubt still lurks. "Do I ever think I'm not qualified?" she says. "All the time."

Not a syndrome

Although often referred to as the 'impostor syndrome', the affliction is not recognized medically as a syndrome — a group of symptoms signifying a disease or the propensity to develop one. In their 1978 paper, Clance and Imes deliberately avoided using the word syndrome. "I didn't want this to be one more way of pathologizing women," says Clance. She and Imes initially believed the affliction affected mainly women, but later research has shown that men fall prey as well. Many people identify themselves as sufferers, but it's a matter of debate as to whether IP is actually a distinct emotional or psychological condition. "A whole lot more has been made of this for commercial purposes than it deserves," says Mayada Akil, medical director of the

outpatient programme in the psychiatry department at Georgetown University Hospital in Washington DC. In recent years, it has drawn significant interest as the subject of books, talk shows and popular-magazine articles. "I think the emotional experience exists, but it is not a disorder or syndrome," says Akil. Rather, she suggests, other underlying maladies such as depression are responsible for severely impeding work. Depression is a true disorder, she says.

But for believers such as Clance and Imes, IP is a real condition that often has serious consequences. It may emotionally choke its sufferers to the point at which it derails their career. "One of my clients had been offered this promotion, but she decided she didn't have the skill set to accept it. She declined," Clance remembers. Although this story has a happy ending — the client's unusually perceptive superior encouraged her to accept the new post and told her she could return to her former post if things didn't work out — many don't, Clance warns.

Recognizing the source of self-doubt: David Poeppel (left) and Pauline Clance.

The search for perfection

Catherine Cardelús, an assistant professor in biology at Colgate University in Hamilton, New York, first noticed the phenomenon as a graduate student working in the rainforests of Costa Rica. Cardelús is physically fit, and her studies involved a great deal of challenging field work, such as climbing trees, and much intensive lab work. "I was thinking, 'I'm not cut out for this. I really can't do this' — even though there was ample evidence to the contrary," she remembers. And the negative, self-sabotaging thoughts didn't stop, they continued through her postdoc and into her full-time positions. "I'm a high achiever and I'm successful, but I've had those moments of waiting for someone to tap me on the shoulder and tell me I don't belong," she says. "Or to tell me I was really lucky to get that paper in that journal." Cardelús has discussed the issue at international science conferences. "It's pervasive," she says. "Absolutely pervasive."

Clance, a professor emeritus at Georgia State University in Atlanta and a clinical psychologist since 1976, calls Murray's and Cardelús's experiences "classic IP". Neither she nor Imes, a psychotherapist and clinical psychologist in Atlanta, can definitively say whether women in science suffer from the problem in greater numbers than those in other fields or than men in any field.

Although Murray and Cardelús say it has not damaged their careers, both admit they haven't published as many papers as some of their male counterparts — perhaps in part because neither can bear to have so much as a typo mar their work. So they review, rewrite, proofread, review, rewrite and proofread. The process can take years. "I sit on my papers," says Cardelús. "I make sure there's not a single typo and if there is, it's a disaster." Indeed, the need for perfectionism is a hallmark of feeling like an impostor, according to Clance and Imes. "They have to be the perfect hostess and mother and wife. They have to be perfect at everything," says Imes.

What causes this emotional self-sabotage? Clance and Imes found that childhood experiences typically begin the cycle. Sufferers were often valued for their intelligence, giving rise to self-doubts and feelings of fraudulence when excellent grades don't materialize in graduate school and, later, when a new postdoc or new job isn't a breeze. "A lot of high achievers grew up in families where they are given approval for achievement but not given much validation for their feelings," Imes explains. "So they grow up thinking their worth or value is tied only to achievement."

New York University neuroscientist David Poeppel struggled for years with impostor feelings that emerged when he was a graduate student at the Massachusetts Institute of Technology and continued to plague him through his postdoc at the University of California, San Francisco. "I was working on research I thought was innovative, but was surrounded by overachievers," he says. "You think, 'I couldn't have done that. How come I'm even here at all?'." It still sometimes raises its unwelcome head. "Exogenous validation doesn't do it," Poeppel says. "Somehow, you can't internalize. You cannot say, 'All right, what I'm doing is serious and is taken seriously and I should be calm about it and just chill'."

Underlying anxiety

So how do the afflicted accomplish anything at all, stymied as they are by self-doubt and the sense that they're fooling everyone around them? There are several steps that may help (see 'Purging that inner critic'). Clance and Imes encourage sufferers to seek therapy. Many of their clients who struggled with IP, they say, also had underlying depression or anxiety that responded to counselling and medication, which also relieved the impostor feelings.

Becoming responsible for others has also helped some sufferers, perhaps because the focus is turned away from themselves. Poeppel says his impostor feelings began to subside after several years as an assistant professor, about eight years ago. When he was no longer competing with classmates and colleagues for grades and accolades and was instead shepherding younger students along their own paths, the self-doubt and terror mostly fell away. "You feel OK about complimenting people on their performance," he says.

"One of my strategies is to reread papers of mine and remind myself, 'Wow, that was great, that was such a good paper,'" says Cardelús, who adds that having her first child also alleviated some of her struggles. Being directly responsible for a new life that she helped create has given her a sense of accomplishment and mitigated her impostor fears. "That's a huge moment in your life and it should empower you," she says. ■

Karen Kaplan is assistant editor of *Naturejobs*.

PURGING THAT INNER CRITIC

Here are some strategies that may help those struggling with impostor phenomenon.

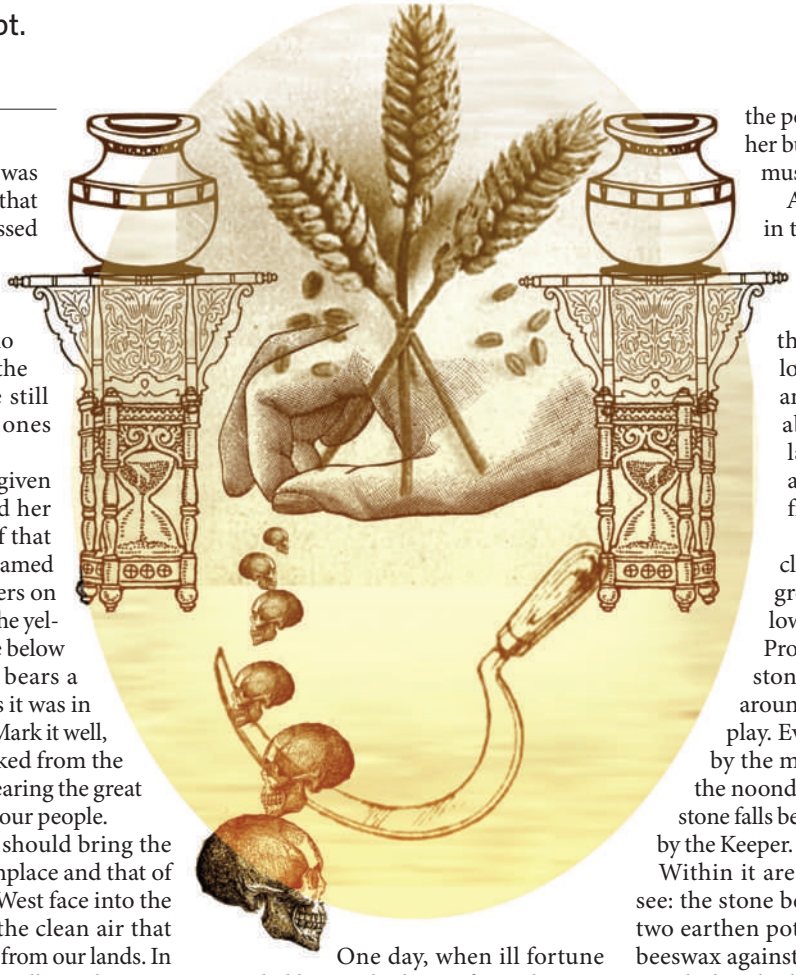
- If you're a student or a postdoctoral fellow, get a supportive, understanding adviser.
- If you're working, do your best to find a supportive, understanding mentor.
- Call on your partner or friends to be supportive and talk you through impostor feelings.
- Hire a tutor or take a class in a topic or area where you think your abilities are weaker. You'll learn what you need — or realize how much you already know.
- Make a list of your strengths. Look back at examples of your own successful work, or positive reviews, and remind yourself of your own accomplishments.
- Accept that some tasks will not be done perfectly.
- Be aware of your language choices. If you find yourself thinking you were 'lucky' to have got a grant or published a paper, focus on what you did to earn it.

K.K.

Final protocol

The seeds of doubt.

John Gilbey



JACEY

He who fell from grace was called the Professor — that is to say, one who professed grand skill and art. The term is now ironic, as his fall was great and its harm greater. She who sought to save us was the Doctor — the title we still give our healers, the ones who give us hope.

We are told that her given name was Heather and her eyes were the colour of that same plant — perhaps named for her — that yet flowers on our cliffs. Her hair was the yellow of the beach you see below us. This central stone bears a rendering of her face as it was in life so many years past. Mark it well, it shows where she walked from the sea, naked and alone, bearing the great Treasure — the hope of our people.

It was meet that she should bring the Treasure here, her birthplace and that of her parents. We of the West face into the wind from the sea — the clean air that helps to drive contagion from our lands. In her mind, the chronicler tells us, this cove and the fields between the mountains and the sea stood as a last redoubt.

The distant great hall where the Professor and Doctor dwelt had many rooms, all given to the study of plants. The Doctor took wayside flowers and grains, and teased from them new strains that gave more bounteously. How noble this study seems to us in our parlous state! But there were also dark ambitions, to which the Professor was privy, which sought blights with which to destroy the food of others. The Professor was also a smoker — a word whose meaning is distant to us. To continue living he had to perform certain rites periodically — outside, as the Law demanded. By breathing the smoke of various herbs he claimed to reach a state of nirvana, but this was achieved at a cruel cost to us all.

The Doctor, as mother of her craft, held the Book of Protocols: powerful writ that tasked her folk with many things — the way of how things must be done, and how they must not. Chief among these oaths was laid the Protocol of Security — which sought to contain both the good and evil within the hall, lest they escape to make untimely demands upon the world.

One day, when ill fortune attended his work, the Professor let vent great fury and stormed from the hall to smoke — still in his work attire. This defied the Protocol and caused the Doctor and her people great woe. She upbraided the Professor with strong words — which he scorned, forcing her to the presence of the Men of Power. Despite her earnest entreaties they spurned her call for action — merely admonishing the Professor as we would chide a child that strays.

Time passed and all seemed well, so the Men of Power smiled and nodded in their familiar way — until troubling news reached them. Soon, the air spoke of a blight sweeping the land — dark chancres covered every grain and those that ate them grew crazed. Men fought with fire — first the sickness and then each other — until the whole country was laid to ruin and starvation. We of this coast were lucky — few came near us, and those that did we dealt with. As our grass and grain withered we turned once again to the sea for fish and the weeds of the shoreline.

Through this terror, amid many horrors and privations, Heather fought her way home. Slipping from the fishing boat that carried her the last miles, she called on the power of the sea to cleanse her burden of the sickness that must surely taint it.

As she lay, mortally ill, in the house of her mother, Heather called to her a friend of her youth — a yeoman who yet worked the land hereabouts. His love for Heather ran deep, and he cried to see her so abused. Yet, there was a last task that they must address — to write the final Protocol.

Now, gather and attend closely — for when you are grown, one of you will follow me as the Keeper of the Protocol. Beyond the central stone you can see the cairn, around which you are wont to play. Every tenth year, counted by the marks you see here, once the noonday shadow of the central stone falls below the cairn it is opened by the Keeper.

Within it are things you have yet to see: the stone bearing the Protocol and two earthen pots, their lids sealed with beeswax against the salt air. Within one pot, the hundred small packets of Treasure lie — formed of a bright metal thinner than our blacksmith can devise, and which can be torn between the fingers. Fifty of these packets are now empty — the grains they carried long dead in the field behind you.

One day, it is reasoned, seeds will be sown that will flourish — the blight being scoured from the land by sun and rain. Certainly, it is said that Heather chose the seeds from the finest she could devise. Perhaps it will be one of you whose hand sows them, as the Protocol dictates. When that day comes we will save and multiply the grain until we have enough to go beyond the mountains, into the grey wasteland beyond, so that man may once more have grain for his food.

What's that? The other pot? Well, on her death the body of Doctor Heather was burned with due ritual in the limekilns at the edge of the beach. That pot contains her ash — a pinch of which is spread to help nourish each crop. Mayhap her beauty and skill will pass into all those who one day eat of her grain! ■

John Gilbey is a gentleman and a scholar, currently Visionary in Residence at the University of Rural England.